DOCUMENT RESUME

ED 088 040 CS 000 989

AUTHOR Bormuth, John R.
TITLE Reading Literacy: Its Definition and Assessment.
PUB DATE 73
NOTE 61p.; Paper prepared for the Committee on Literacy of the National Academy of Education; reprinted from Reading Research Quarterly, 1973-1974

EDRS PRICE MF-$0.75 HC-$3.15
DESCRIPTORS Cloze Procedure; *Literacy; Reading; *Reading Ability; Reading Achievement; *Reading Comprehension; Reading Development; *Reading Research; *Reading Skills

ABSTRACT

The purposes of this article are to analyze the concept of literacy in order to identify measurement problems associated with specifying each of these parameters, and to describe literacy assessment procedures now available for dealing with measurement problems. The principal focus of the paper is on the development of models for identifying performance criteria that can serve as the goal of instructional programs and of the research and development programs that lead to them. The five parameters discussed are the classes of literacy behaviors, the level of performance that serves as the criterion of literate performance, the kinds of reading tasks on which the behaviors are tested, the proportion of the reading tasks that serves as the criterion of literacy on some corpus of reading tasks, and certain characteristics of the prople tested, such as the levels of aptitude and perseverance represented within it. (Author/WR)

ED 088040

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

*Reading literacy: its definition and assessment**

JOHN R. BORMUTH
University of Chicago

Reprinted from
READING RESEARCH QUARTERLY, Volume IX, Number 1
1973-1974

Published by the
International Reading Association, Newark, Delaware
Copyright 1973 by the International Reading Association

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
International Reading Association
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

CS 000 989

* This paper was originally prepared for the Committee on Literacy of the National Academy of Education. It will also appear in a collection to be published by the National Academy under the title *Toward a Literate Society*.

*Reading literacy: its definition and assessment**

JOHN R. BORMUTH
University of Chicago

THIS ARTICLE HAS 3 PURPOSES: 1) to analyze the concept of literacy for the purpose of identifying the parameters that must be specified in literacy definitions, 2) to identify measurement problems associated with specifying each of these parameters, and 3) to describe literacy assessment procedures currently available for dealing with these measurement problems. The principal focus of the paper is on the development of models for identifying performance criteria that can serve as the goal of instructional programs and of the research and development programs that lead to them. The 5 parameters discussed here are a) the classes of literacy behaviors, b) the level of performance that serves as the criterion of literate performance, c) the kinds of reading tasks on which the behaviors are tested, d) the proportion of the reading tasks that serves as the criterion of literacy on some corpus of reading tasks, and e) certain characteristics of the people tested, such as the levels of aptitude and perseverance represented within it.

L'Aptitude à la lecture: définition et évaluation

CETTE ETUDE A 3 BUTS; 1) analyser le concept de l'aptitude à la lecture afin de constater les paramètres qui doivent être spécifiées dans toute définition de cette aptitude; 2) diagnostiquer les problèmes de mesurage associés à la spécification de ces paramètres; et 3) décrire les procédés d'évaluation de l'aptitude à la lecture que l'on emploie de nos jours afin de résoudre ces problèmes de mesurage. Cette étude se concentre principalement sur la découverte de modèles qui meneraient à l'identification de critères de performance. Ces critères pourraient alors servir

* This paper was originally prepared for the Committee on Literacy of the National Academy of Education. It will also appear in a collection to be published by the National Academy under the title *Toward a Literate Society*.

comme but ultime des programmes d'instruction de même que des programmes de recherches qui aboutiraient à l'établissement de ces programmes d'instruction. Les 5 paramètres discutées sont: a) les catégories de comportement dans l'aptitude à la lecture; b) le niveau de performance qui sert de critère à l'accomplissement de la lecture; c) les diverses tâches d'après lesquelles le comportement est mis à l'épreuve; d) la proportion de tâches acquises parmi un corpus de tâches qui pourrait servir comme critère de l'aptitude à la lecture; et e) certaines caractéristiques des individus mis à l'épreuve comme, par exemple, leur niveaux d'aptitude et de persévérance.

Capacidad de leer: su definición y determinación

ESTE ARTICULO TIENE 3 PROPOSITOS: 1) analizar el concepto de "capacidad de leer", con el proposito de identificar los parámetros que deben ser especificados en las definiciones de "capacidad de leer", 2) identificar los problemas de medición asociados a la especificación de cada uno de estos parámetros, y 3) describir los procedimientos para determinar la "capacidad de leer" actualmente disponibles para poder tratar estos problemas de medición. El principal enfoque del artículo es en el desarrollo de modelos para identificar los criterios de desempeño que pueden servir de objetivo en los programas de instrucción y en los programas de investigación y desarrollo que conducen a ellos. Los 5 parámetros tratados son a) las clases de comportamientos en la "capacidad de leer", b) el nivel de desempeño que sirve de criterio en el desempeño de la "capacidad de leer", c) los tipos de tareas de lectura mediante los cuales se prueban los comportamientos, d) la proporción de las tareas de lectura que sirve de criterio de la "capacidad de leer" en algunos cuerpos de tareas de lectura, y e) ciertas características de la gente examinada, tales como los niveles de aptitud y perseverancia representados en él.

Literacy may be defined broadly as being able to respond appropriately to written language; in this sense, it is one of man's most valued skills. Man has used writing to record, accumulate, and store his knowledge in an easily used form. Because those who were literate have been able to overcome the barriers that time and space throw in the way of communication, some have been able to master and apply technical information and thereby achieve unprecedented material prosperity. Some have been able to master and apply social and political knowledge to secure personal and political liberties for themselves. And some have been able to enlarge their perspective and satisfy their aesthetic desires through literature.

Literacy is an undeniably great benefit, but only to the literate. During the past century, nearly every movement that has sought to better man's lot has given a prominent place in its program to making him literate. And all of these programs have eventually encountered the same problem. When their proponents descend from the rarefied stratosphere of rhetoric and attempt to implement their programs, they must ask the complex question: what does it mean to be literate? None of the many approaches which have tried to answer it has provided more than a narrowly limited answer. This paper will once again address the issue, not with the naive aim of being able to answer the question with a single deft stroke but rather with the humble hope of being able to identify most of the major parameters of the answer and of being able to suggest what general form the ultimate answer might take and how it might be arrived at.

Conceptions of literacy

Let us begin by examining some of the earlier efforts to define and assess literacy. We hear claims that there remain large numbers of illiterate people in the United States, a nation that has experienced several generations of free and compulsory public education. The late James E. Allen, Jr. former US Commissioner of Education, cited these figures (1969):

1] One out of every 4 students nation-wide has significant reading deficiencies.
2] In large city school systems up to half of the students read below expectation.
3] There are more than 3 million illiterates in our adult population.

4] About half of the unemployed youth, ages 16-21, are functionally illiterate.
5] Three-quarters of the juvenile offenders in New York are 2 or more years retarded in reading.
6] In a recent US Armed Forces program called "Project 100,000," 68.2 per cent of the young men fell below grade 7 in reading and academic ability.

If these statements indicate that there are large numbers of illiterate citizens within the United States, they may also be taken as evidence that educational institutions have failed tragically to achieve one of our most deeply-rooted aims—that all men should have equal opportunities to develop and attain their ambitions. Such reasoning seems to be what prompted Commissioner Allen and others to advocate massive research and development programs aimed at developing literacy instruction that could remedy the problem.

Need for better literacy assessment procedures

Certainly, if the illiteracy level is so high, such programs would seem urgently needed and large expenditures of public moneys justified. Unfortunately, however, it is impossible to put much faith in these or in any other literacy statistics currently available; for none of them is based either on a careful analysis of the concept of literacy itself or on suitable methods of measurement. It may be worthwhile to examine some of the more commonly used procedures for assessing literacy and to briefly describe some data indicating that the literacy problem may be far more serious than these procedures would lead us to believe.

Functional literacy. The Bureau of the Census attempts to assess the literacy of the population by tabulating the number of people 14 years of age or over who have not completed 6 years of school. This constitutes the criterion for what is called *functional literacy.* In order to accept figures based on this criterion, it is necessary to make several dubious assumptions; but just one needs to be examined here.

There is no evidence to support the assumption that 6 years of schooling are sufficient to raise all students' abilities to the point where they can deal competently with ordinary reading tasks. One study (Bormuth, 1969c) found evidence that it is probably false. A fairly representative sample of 8 articles was drawn from news publi-

cations, a cloze readability test was made over each,[1] and these tests were administered to students in grades 3 through 12. These were children from middle-class homes in a residential suburb of a large Midwestern city. The average percentage of students who were able to answer at least 35 per cent of the cloze questions on the tests was calculated. On the average article, only 33 per cent of the students in grade 6 and only 65 per cent of those in grade 12 reached this criterion.

In other studies (Bormuth, 1971) it had been shown that students who are unable to answer at least 35 per cent of the items in a cloze readability test can gain little or no information from materials at that level of difficulty. Consequently, there seems to be little basis for claiming that a person completing 6 years of school is literate. Even graduation from high school does not appear to be a very certain criterion of literacy. The fact is that the number of years a person has been in school is a very poor index of his ability to read, for within any grade level it is common to observe very wide variations in the reading abilities of the students. But even if grade level were an accurate index of literacy, the small amount of evidence that is now available would indicate that the grade 6 criterion is far too low and that the illiteracy rate is probably much higher than the Census Bureau would lead us to believe.

Achievement level. Others have tried to get around this problem by using, in some way, a person's achievement grade level instead of the years he has spent in school. This is a number found by giving a reading test of some sort to students who are in various grades in school, say at the 5.2 grade level. Their mean test scores are calculated, and thereafter students who get a score equal to the mean of this group are assigned to a grade level at which, on the average, students are able to answer that number of questions on that test.

But, again, it is hard to tell what these grade level scores mean. Commissioner Allen cited a study reporting that 68.2 per cent of the men in an Armed Forces study had grade scores of less than 7.0. It is impossible to say what this means. Can students with 7.0 scores read newspapers, college textbooks, or even the text in comic books competently? A grade level score does not provide us with any information on just what kinds of real-world reading tasks a person can per-

1. The cloze readability procedure will be discussed in some detail later in this paper.

form competently. Consequently, we learn little about the level of illiteracy in the population when grade level scores are used to tell us either that some proportion of the population falls below a certain grade level or that some proportion is 2 grade level years below their current grade in school.

Data from a study already cited (Bormuth, 1969c) shed some light on the matter. By performing a series of regressions between scores on cloze readability tests made from each of several newspaper articles and a test that gave grade level scores, it was possible to calculate that the grade level score of the average person who answered 35 per cent of the items on the cloze readability test was 10.5, indicating that the average person is literate with respect to half of the newspaper articles only after 10.5 years in school. This indicates that the study cited by Allen employed a criterion that was far too low and also that the illiteracy rate may be very much higher than estimated by that study.

Grade level expectancy. The third major way in which people attempt to assess literacy is by using the *expectancy* concept. According to this, a person has some level of aptitude for learning reading. This is usually measured by a verbal aptitude or verbal intelligence test score that is converted to a grade level score. This grade level score is said to be the person's reading expectancy, meaning that if he were working "up to his capacity," he would probably get a similar grade level score when he is given a reading achievement test. Hence, a person whose achievement score is, say, 2 years below his expectancy score does not seem to be profiting very well from his instruction.

It is possible to cite several strong statistical reasons why the studies that have reported these kinds of data in the past must be viewed with suspicion. But these can be put aside for an even stronger reason: the expectancy score is based on 2 grade level scores, neither of which tells anything about whether a person can perform competently on real-world reading tasks. In passing, it may also be worth mentioning that if every student were working exactly up to his capacity and if the tests now used to measure capacity and achievement were slightly unreliable (as all tests are), then exactly half of the students would appear to be working below capacity at all times—a phenomenon Allen cited as evidence of extensive illiteracy in large city schools, but one that may be merely an artifact of the random variation in test scores.

Proportion below grade level. One occasionally reads a report that such and such a percentage of the students in some school system fall below grade level and are, therefore, destined for a life of illiteracy. If this prediction were true, it would not be because the children fell below grade level. The grade level scores represent nothing more than the mean scores of the students in a given grade level. Hence, if the achievement test were well-made and fairly recent, we could always say with very good accuracy that half of all children are always below grade level at all times. But, because we are dealing only with grade level scores, we still have no idea whether all, some, or even none of those children whose scores fell below grade level can respond competently to real-world reading tasks.

These remarks should not be interpreted as criticism of Allen or of anyone else who has attempted to deal with the literacy problem at the policy-making level. He and other men of good-will sense that something is amiss in literacy training—that large numbers of people probably never reach a level of reading ability sufficient to cope with even the common reading tasks confronting them daily. In order to rally the support needed to remedy the problem, these men require evidence. It is extremely unfortunate that there is as yet no adequate evidence to place in their hands. But the fact is that we have not yet analyzed exactly what is meant by literacy and then devised appropriate methods for measuring it.

Nature of literacy

The term *literate* may be used to refer to a number of different kinds of behavior, ranging from the ability to employ basic reading or writing skills to the knowledge of some body of literature. The term will be used here to refer to the ability to respond competently to real-world reading tasks. To define the term further, however, requires that we give detailed specifications of 5 parameters: 1) the behaviors we wish to observe, 2) the criterion level of performance we expect a literate person to demonstrate on tests of those behaviors, 3) the kinds of materials on which we test the behaviors, 4) the criterion proportion of the reading tasks on which the person must exhibit a literate level of behavior, and finally 5) certain characteristics of the person tested, such as his aptitude and practical needs (goals).

Comprehensive and fragmentary programs

The 2 fundamentally different approaches taken to the design of literacy programs can be referred to as *fragmentary* and *comprehensive*. Accept, for the moment, the proposition that a person is literate if he can obtain all the information he needs from the materials he needs to read. If we view literacy in this way, we can see that there are 2 major determinants of a person's literacy—1) how many comprehension skills are required by the materials that he needs to read and 2) how many of those skills has he mastered? Up to the present, virtually all literacy programs have been fragmentary. In their conception, planning, and execution, they have attempted merely to manipulate only one of these 2 determinants. And these programs have either ignored the other determinant or regarded it as unchangeable. A comprehensive literacy program takes both determinants into account. Such a program seems feasible within our current social structure and would probably be more effective and economical than fragmentary programs. Moreover, I can see no way to define and assess literacy meaningfully unless we take both determinants into account.

In order to make these matters clear, let us examine a brief analysis of the literacy system as it operates in our society. The primary purpose of literacy is to enable a person to gain information from material. This information produces effects on his behavior that are considered sufficiently desirable to society to warrant its paying for the individual's instruction to master the literacy behaviors. And the effects are considered sufficiently desirable to the individual to warrant his spending considerable amounts of time and effort acquiring the literacy skills. This creates a demand for materials containing information of various sorts; and the publisher's job is to determine what kinds of information are needed, to arrange to have that information prepared in written form, to edit this material into a form that meets the needs of the consumer, and then to print and distribute it. The publisher's reward is great enough that he is willing to use whatever reasonable means are available to edit and tailor the readability of the materials so that they require just those literacy skills that the consumers are most likely to possess. Conversely, society's rewards are great enough that it attempts to instruct its members in whatever literacy skills the materials may require.

A person can be considered literate in this system when he can get the information he needs from the materials that he needs to read. And this is true regardless of whether we view the matter from his, the publisher's, or society's standpoint. Hence, a person may be regarded as literate or illiterate only with respect to a particular reading task; and his status relative to that material may be altered both by giving him instruction in literacy skills and by altering the materials so that the literacy skills that they require match the ones he has learned.

It seems fairly clear that the primary things required in order to mount a comprehensive literacy program are the necessary technologies required to assess and manipulate the readability of materials and to teach and assess people's mastery of the literacy skills. In the system outlined, the motivations of the individual, the society, and the publisher would assure that they would adopt the techniques. Some coordination would be necessary to insure that the materials were tailored to the skills being taught and that the curriculum of this instruction taught all of the skills that are essential for transmitting information. Moreover, this coordination could lead to considerable economy in both the instructional and publishing activities. The number of skills taught could be limited and tailored to fit the personal needs of individuals, effecting a savings in instructional costs. Similarly, the materials could be tailored to fit the skills of their intended audiences and thereby increase the market and rewards for the publisher.

Moreover, there appears to be no sensible way to define and assess literacy if we conceive of literacy in a fragmentary way. As this discussion progresses, it will become evident that there are a huge number of skills that we could consider literacy skills, probably far more than we could afford to learn. However, there is no way to select among these skills unless we take into account the materials in which these skills are required and the usefulness of those materials to society and to various types of individuals.

Past programs have been conceived and organized primarily as fragmentary programs out of administrative necessity. Because of the tradition of local control of schools in the United States, literacy programs have had to be carried on at that level, utilizing only local resources and affecting only a negligible proportion of the total population. Such a program could not be designed to teach just certain

literacy skills, for doing so might have made its students illiterate with respect to a large proportion of the materials being published. And, since only a small proportion of the population was affected by a particular school's curriculum, publishers could not afford to prepare materials especially for them. At the present time, however, there have been many precedents for obtaining adequate funding and administrative coordination from foundations and government to design programs that would have the administrative breadth to affect the literacy of nearly all students in the United States. The work of the so-called modern mathematics and science programs provide graphic examples of what is possible. Under such circumstances, publishers might be willing—perhaps even eager—to cooperate.

But had adequate funding and administrative coordination been available at a much earlier date, it is doubtful that a comprehensive program could have had much, if any, effect. There simply was not an adequate scientific base on which to build the necessary technology. Reading instruction and readability were practiced as crafts, whose effectiveness depended heavily on the experience and intuitions of the practitioners, rather than as technologies, which could be employed to produce predictable results. We could not identify in any reasonably acceptable way, for example, what skills were involved in literacy, what features of language were involved in those skills, or how the language features and their associate literacy skills influenced the difficulty of materials. Nor did anyone know quite how to go about finding out about such things. While we still remain largely ignorant of the nature of literacy skills, psychologists, linguists, and psycholinguists have, in the course of the past few decades, built a scientific base that seems to be at least adequate for the effective study of these matters.

Kinds of behaviors

The first component of a literacy definition is a set of statements describing the kinds of behaviors a person must be able to exhibit in order to be classified as literate. Pertinent to this are discussions of a) the range of behaviors that must be considered (though not necessarily included) whenever a definition of literacy is formulated; b) the need to limit the range of behaviors included in a particular definition intended for practical use in research, development, or instructional programs and the more important criteria for select-

ing those behaviors to be included; and c) some of the major measurement problems involved in designing tests for assessing literacy.

Range of behaviors involved in literacy

At first glance it would not seem to be a particularly difficult task to say just what behaviors are implied by the term *literacy*. To say a person is literate seems to claim that he can perform some set of reading tasks competently. So all one would have to do to arrive at the sought-after literacy behaviors is to analyze those tasks to see what behaviors they required. But this first glance is deceptive, for this problem is closely associated with another problem containing several complexities that have led to heated and emotionally charged controversies. These controversies arose out of the question of whether the reading act involves just the *word recognition behaviors*—those skills involved in decoding written words into spoken words—or whether it *also* includes such behaviors as comprehending that language, critically evaluating its truth and relevance, appreciating its aesthetic qualities, and so on. When this problem is properly analyzed, it reduces not to an either/or question but merely to a series of questions about priorities which can be rather easily (but not painlessly) resolved on the basis of values shared by the protagonists on both sides of the argument.

Controversy. Although this controversy has existed for a very long time in the area of reading instruction, it surfaced and became a full-blown public controversy with the publication of *Why Johnny Can't Read* by Rudolph Flesch (1955). Flesch noted that substantial numbers of children were unable to perform competently even the most rudimentary reading behavior—decoding written words into spoken words; and he attributed this fact not only to a lack of phonics content in the reading curricula used in schools but also to the presence of a considerable amount of instruction designed to teach students the higher level skills commonly referred to as *comprehension*, *critical reading*, *literary appreciation*, and the like. It was not his contention that these were unimportant skills to learn. Rather, it seemed to be his belief that these higher level skills were of secondary priority in the sense that they could not be learned until the decoding skills had been mastered and that their early introduction into the curriculum interfered with word recognition instruction by diverting energy away from the acquisition of decoding skills.

But a confused controversy has continued in other forms among psychologists, linguists, and educators, centering, among other things, on the issue of whether reading curricula should include instruction in the higher level skills. Psychologists and linguists have argued that reading can be conceptualized as only those skills uniquely involved in decoding written language into spoken language and that everything else in the reading curriculum does not really teach reading skills at all, but rather something often vaguely lumped together and labeled *thinking skills*. A number of others, mostly reading specialists, have taken the position that the reading act could not really be broken up in this way. They argue that there is an underlying continuity in the reading act, that such a distinction is arbitrary, and that omitting instruction in the higher level skills would cripple children's potential for performing useful reading tasks.

This argument would have long since evaporated had the protagonists begun by addressing themselves to the same issue. The group that wishes to define reading as being coterminous with the decoding skills has included largely scientists in linguistics and psychology. To them identifying reading behaviors primarily involves breaking them down into small classes so that they can plan and carry out manageable scientific analyses. A scientist simply cannot perform useful theoretical work until he has obtained rigorously defined classes of phenomena to study; thus for their purposes, these scholars were absolutely correct to place the decoding skills in a class by themselves in order to provide a fairly natural and manageable phenomenon that can be analyzed from existing linguistic and psychological theory.

On the other side of the argument, one finds mainly the specialists in reading instruction. Their objective is to provide a complete system of behaviors to permit students to cope effectively with the reading tasks encountered in the real world. When the specialists analyze these real-world reading tasks, they see that the students must learn not just the decoding behaviors but also the higher level skills, which their opponents *seem* to oppose teaching. The reading specialist then labels all of these skills *reading skills* but without making it clear to others that he uses this label merely as a convenient method of referring to *anything* that is taught during the period labeled *reading* in the schedules which appear in curriculum guides. To

him, the label refers to instruction in how to turn a book's pages, in how to find and read page numbers, in decoding skills, and in any other behavior that he thinks a) is functional in coping with real-world reading tasks and b) can be more conveniently taught during that time period than, say, during the period labeled mathematics. Both groups, then, are apparently led into thinking that they are talking about the same thing because they both wish to use the same label. And since each side has developed its definition through careful reasoning, it seems to feel the need to jealously defend its usage against anything that appears to be a rival definition. Yet since each definition was designed to serve quite different purposes, they are in no way rivals.

Seen in this light, the problem of choosing the behaviors to be included in a definition of *literacy* is not a problem of identifying what is truly a reading behavior. Rather, the selection of the behaviors to be included in a given definition depends upon the consideration of the purpose that definition is to serve. If its purpose is purely scientific, then the criteria of conceptual and theoretical tractability seem appropriate for identifying those behaviors. But if the definition is to serve as the statement of the objectives of an instructional program that purports to develop a system of behaviors having utility in the real world, then it is appropriate to apply stringent social, political, cultural, and economic criteria in the selection of those behaviors.

What is important to note at this point is that there is no *true* definition of literacy. Rather each definition must be designed for the purpose to which it is to be put, and its correctness may be judged only in terms of how well it serves that purpose. Thus, when a definition of literacy is being developed, it would seem rational to state clearly the purpose of that definition, to derive from this statement a set of criteria for selecting and excluding behaviors, and then to select behaviors using these criteria. It seems likely that had rational procedures of this sort been followed in the earlier formulations of the concept of literacy, we might have been spared much pointless and often destructive controversy.

Taxonomy of literacy behaviors. Much effort has gone into the matter of identifying the behaviors a person must have in order to deal with a variety of reading tasks. Collecting these taxonomies has been largely performed by curriculum specialists in reading but

much of the content itself has been contributed by analyses in the disciplines of psychology; in manuscript criticism in the study of history, linguistics, and library sciences; and in a number of other areas.[2]

First, there are the *decoding* behaviors, which enable a person to map letters, letter groups and patterns, and typographical features of print onto oral language units. Normally this includes the phonics behaviors, which map the smaller graphological units onto language sounds; the word structure behaviors, which map whole syllables and affixes as units onto their corresponding sounds; the sight recognition behaviors, which map whole words onto their corresponding sounds; the context recognition behaviors, which utilize the context surrounding a word to map the word onto its sounds; and the dictionary behaviors, which enable a person to locate and pronounce a word from its entry in the dictionary.

Second, there are the *literal comprehension* behaviors, which enable a person to learn the information explicitly signalled in a reading task. This normally includes the vocabulary meaning behaviors, which enable a person to assign the correct meanings to words in their contexts; the sentence comprehension skills, which enable a person to combine the meanings of words in sentences according to patterns conforming to the syntax of the sentences; the anaphora comprehension behaviors, which enable a person to identify the recurrences of concepts in a reading task so that the appropriate concepts are modified when they reoccur in sentences; and the discourse comprehension behaviors, which enable a person to combine the meanings of sentences in a passage according to patterns signalled by the discourse syntax of a reading task.

The remaining classes of behaviors have generally been less well analyzed than the 2 just named. The third might be described as the *inference behaviors*, which enable a person to derive information not explicitly signalled by the reading task. These behaviors might be described impressionistically as those that occur when a person "reads between the lines" or somewhat more formally as being logic-like processes in which statements in a text might be substi-

2. At this point it would be inappropriate to attempt either exhaustive listings or precise definitions of these areas of behavior. More extensive listings may be obtained from other sources (such as Betts, 1954; Bond and Tinker, 1967; or Harris, 1962). And the problem of defining complex cognitive behaviors such as these will receive separate discussion in this article. The brief discussion presented here is provided merely to give the reader a general impression of the range of behaviors that must be considered for inclusion when a definition of literacy is being developed.

tuted into logical algorithms and true sentences not in the text computed by using predicate calculus.

The fourth set of behaviors are generally called the *critical reading skills*, and they conform roughly to the procedures known as manuscript criticism in the study of history. They consist of applying tests of the consistency of the logic of a text, verifying its factual claims, verifying the authority of the writer, and detecting and evaluating propaganda devices.

The fifth set are the *aesthetic appreciation behaviors*. These are difficult to characterize because they are typically discussed in terms that do not readily lend themselves to behavioral analyses, including phrases such as *detecting the tone and mood of the story, seeing the deeper meanings, detecting the pacing or rhythm of the prose*, and so on. This set of behaviors seems to be largely appropriate for just those reading tasks that have aesthetic pretentions.

The sixth set of behaviors have been traditionally known as the *reading flexibility skills*. They are the behaviors that enable a person to speed up or slow down his reading, depending on the nature of the task. They also enable a person to focus on just the parts of the text containing the types of informaion tested by some set of questions or described in some set of instructions, and to switch these attentional behaviors to conform to a wide variety of such instructions. More recently, this set of behaviors have come to be known as mathemagenic behaviors (see Rothkopf, 1966).

The seventh and final category comprises the *study skills*, which include an assortment of behaviors that enable a person to use various reference devices to locate information and then to judge its relevance to some problem. This category also includes behaviors that enable a person to interpret special devices for presenting information, such as maps, graphs, outlines, charts, diagrams, and the like.

Obviously a complete listing of all the behaviors implied by these 7 categories would constitute a work of its own. It should be noted, also, that other classes of behaviors could be added—the primitive reading readiness behaviors, such as those studied by Gibson (1970), for example. However, these rather brief descriptions should be sufficient to enable the reader to get some sense of the full range of behaviors that are included in at least some instructional programs labeled as *literacy* or *reading programs*.

Limiting a definition

In the broadest sense of the word, *literacy* is the ability to exhibit all of the behaviors a person needs in order to respond appropriately to all possible reading tasks. However, it is unlikely that a definition of literacy that specified all of these behaviors would have much utility. *a definition of literacy*, as that phrase is used here, represents a detailed and explicit statement of the goal of a research, development, or instructional program; and all such programs must contend with limitations on funds, time, adequacy of scientific knowledge, access to skilled personnel, and so on. And they must state a reasonably believable goal in the first place even to be granted the use of any resources at all. As a result, they invariably face the need to limit the scope of their goal statements.

One convenient and often necessary way to limit the definition is by including in it only some of the behaviors normally regarded as literacy behaviors. However, this must be done with considerable care in order to avoid serious mistakes. If certain scientific considerations are ignored, for example, the definition may only appear to be sufficiently limited to be useful when in fact it may implicitly commit the program to an impossibly large task. Or, if the definition includes only socially trivial behaviors, the program may fail to win either the financial or scientific support essential for its success. Hence, the matter of selecting behaviors to include in a definition deserves some examination.

Utility. Selecting and validating educational objectives involves problems peculiar to reading instruction. The first has already been discussed in another context. This is the problem that either reading behavior can be viewed as a phenomenon that can be studied usefully to make scientific contributions to basic linguistics, psychology, history, and other areas of study, or it can be regarded as a system of behaviors having considerable economic, social, cultural, and political value both to the individual who has learned them and to the society of which he is a part. While from many points of view this coincidence that reading behaviors have value in both respects may be a happy one, it also occasions some confusion and controversy.

For example, one psychologist (Gibson, 1970) has been conducting an interesting series of investigations of how children learn to recognize printed letters, and she was awarded special recognition

by her fellow psychologists for her contributions to the understanding of the *reading process*. This has occasioned a considerable amount of wonder among educational psychologists, who regard the work as trivial on the grounds that the processes she was analyzing have seldom been the source of much difficulty in instruction. So if the results of all of this research and all other research of the same type were to be applied conscientiously to the design of reading instruction, it would result in almost no improvement in the rate or degree of children's mastery of reading behaviors.

The important point to note here is not whether the academic or the educational psychologist is correct, since in a certain limited sense, both are. Two different value systems can be and have been applied to this single set of reading or literacy behaviors, with the result that the final judgments were quite different depending on which value system was applied. And the same is true of most of the other literacy behaviors. For example, the historian would undoubtedly place a high value on research that contributed to a better understanding of the so-called critical reading behaviors because of the vital role those behaviors play in the development of his theories, or the specialist in literature would undoubtedly place a high value on the analysis of aesthetic responses to literature; yet in the context of instruction these 2 classes of behavior would be assigned considerably different values. Again, different values can be applied to the same literacy behavior, because each behavior functions differently in different areas of activity. Hence, one can identify and include a behavior in a definition of literacy on the basis of its utility, but unless the purpose of the definition and the criteria used for selecting and rejecting behaviors have been made explicit, one cannot do so without a considerable risk of creating confusion.

This is not to say, however, that scholars from academic disciplines have nothing to say about the utility of behaviors for instructional purposes. Quite the contrary, they often have an excellent grasp of how the literacy behaviors with which they are concerned function in real-world reading tasks. The historian, for example, would likely be quite critical of a program that omitted instruction in the critical reading behaviors. He would point out that such a program would produce a population of credulous dolts who could be counted on to learn and believe almost anything they read but who would be continually subject to the manipulation of demagogues.

Finally, when a definition is used to identify the goals of an instructional program, not only must whole classes of literacy behaviors be selected on the basis of economic, social, cultural, and political criteria but also the specific behaviors within each class must be subjected to criteria of utility. For example, some phonics rules apply with very high frequency in commonly encountered words, and so they would generally be regarded as having high social utility. Other rules apply in only one or 2 words and those words occur rarely in English, so these rules are judged to be of low utility.

Hierarchical entrainment of behaviors. Since the cognitive processes underlying reading behaviors are not directly observable, their relationships are not always immediately apparent and the results can have serious consequences. One of these consequences is that, even though the literacy definition specifies that only one set of behaviors will be taught in an instructional program, it may in fact prove to be necessary to teach many additional related cognitive behaviors before acceptable performance on the target behaviors can be obtained.

Such behaviors are said to be hierarchically related (Gagne, 1965). The simplest case of a behavioral hierarchy may be represented by the diagram shown as a → b. Here the letters *a* and *b* represent 2 behaviors in which behavior *b* is the more complex of the 2 and depends upon behavior *a*. An example of a hierarchy of this sort might be knowing the phoneme corresponding to the letter *f*, which would correspond to behavior *a*, and being able to assign a correct pronunciation to the nonsense syllable *FOD*. The latter behavior, of course, depends upon or entrains behavior *a* but also involves unique components. It follows, then, that behavior *a* must be mastered before *b*. A somewhat similar relationship can hold between classes of behavior. These hierarchies are symbolized with capital letters as shown by A → B. In this case, every behavior in class *B* depends upon at least one behavior in class *A*. An example of this kind of hierarchy is that the behaviors of assigning the meaning to printed words depends upon the behaviors in which sounds are assigned to printed words.

If a literacy definition lists a complex class of behaviors, it is implicitly listing the simpler behaviors entrained by that complex behavior. This fact presents a potentially serious problem when literacy definitions are developed for use in instructional programs in

reading. These hierarchic relationships remain only partially understood, and so it is unclear just what may be entrained in complex behaviors like the critical reading skills or the aesthetic appreciation behaviors. It is possible that, when they are subjected to careful analysis, they might prove to be quite simple and easily taught. On the other hand, it is also possible that they could turn out to be extremely complex so that a definition that included these behaviors might implicitly commit the program for which it serves as a goal statement to a course that is quite beyond the resources allocated to that program.

Interactions among behavioral classes. There are very good reasons to doubt that it is possible to draw sharp distinctions between classes of behaviors that are hierarchically related to each other. In those processes that have been carefully studied, we seem to find hierarchic relationships running in both directions. The main evidence of this is that there is no set of decoding behaviors that, taken by themselves, are sufficient to permit the pronunciation of all the words a person is likely to encounter. The phonics skills, for example, have often been offered as the word pronunciation method *par excellence*. And, indeed, they probably do represent one of the most useful sets of behaviors one can employ to pronounce words.

It is now clear, however, that the phonics skills cannot be employed to pronounce many words unless those skills are coupled with certain of the literal comprehension skills. An obvious example is the printed word *read* in the clauses *they read it yesterday* and *they read it daily*, where one cannot apply the appropriate phonics rule to the vowel letters until he has read the rest of the sentence and comprehended it well enough to determine the tense of the verb *read*. The printed word *lead* in the sentences *they lead their dogs* and *it is made of lead* presents a somewhat different situation in which the application of the correct phonics rule to the vowel letters depends only on the person's having assigned the word to the appropriate part of speech—a process that is thought to be an essential component of the language comprehension processes (Osgood, 1963).

Venezky (1967) has investigated this matter in some detail and has shown that there is a class of words to which the phonics rules cannot be applied directly but only after the word has been assigned to a part-of-speech category. The printed words *suspect, relay, imprint*, and *permit*, for examples, are pronounced differently,

depending on whether they are employed as verbs or nouns. Although this constitutes a fairly small class of words, Goodman (1969) has been able to provide a substantial amount of evidence to show that the comprehension behaviors are employed extensively by children to aid them in the word recognition processes.

Ordinarily, the reading comprehension behaviors are analyzed as hierarchically entraining the word decoding behaviors, and it was pointed out that relationships of this kind must be taken into account in selecting behaviors to be included in a literacy definition. The foregoing discussion demonstrates that a *reverse hierarchy* of a sort operates to connect the same 2 sets of behaviors. Furthermore, it seems likely that these 2-way hierarchies may prevail among a number of classes of behaviors. Research by linguists shows that while language at a higher level of analysis, say the morphological level, is built up out of units from a lower level of analysis, the phonological level, many of the phenomena at the lower levels cannot be explained except in terms of the theory employed at the higher levels.

Measuring literacy behaviors

Deciding what types of behaviors one ought to expect of a literate person presents one type of problem, but deciding how those behaviors should be observed and measured presents problems of a completely different order. The former is primarily a matter of social policy-making in which one decides what social, political, cultural, and economic values are affected by each class of literacy behaviors; weights each class of behaviors according to the weight given each value affect; and then includes in the definition of literacy as many of the most valued behaviors as practical circumstances will justify. Measuring and observing those behaviors, on the other hand, is a scientific and technical problem that involves constructing a theory of the processes underlying those behaviors and then identifying test tasks that can be performed by all, and only all, persons who have actually acquired those behaviors. Consequently, discussion of such testing must deal primarily with the logical and scientific issues involved in testing literacy behaviors.

The argument pursued here has this general form: First, it is economically and logically desirable to use verbal questions as the primary mode of testing literacy behaviors. Second, traditional meth-

ods of deriving verbal test questions are primitive (they do not provide us with the explicit rationale that seems essential for tests that operationally represent research and development programs, especially for those programs which have either serious scientific pretentions or a responsibility for accounting for the effectiveness with which their funds were used). Finally, techniques have become available for developing adequate rationalized literacy tests.[3]

Although it is necessary to restrict our consideration to the problems of testing the more complex literacy behaviors, omitting the word decoding behaviors and the study skills, the arguments presented apply self-evidently and with equal force to the areas eliminated. However, the problems involved in testing the more complex literacy behaviors are much more complicated and have only recently been subjected to analyses that are scientifically adequate, making it more important to focus specifically on them.

Necessity of observing only overt behaviors

In discussions of literacy assessment, as in most discussions of the operations involved in testing cognitive processes, it seems necessary to begin with an *apologia* of 2 rather elementary but very important facts about testing. The first is an explanation of the function of a test item or task, and the second is an explanation of the problems presented by the necessity to observe only overt behaviors.

Function of the test item. Literacy behaviors, like nearly all cognitive behaviors, are not just a set of overt and stereotyped behaviors that a person repeats over and over in nearly identical form, like turning a key in a lock or throwing a ball. People simply are not expected to read the same passage over and over. And when they read, the behavior of major importance is not even observable directly. Rather, a person is expected to exhibit literacy behaviors in response to passages he has never seen before. Thus, a person is literate only when he has learned and can apply a set of *mental processes* that enable him to respond with the appropriate set of behaviors to passages that are new to him.

But a mental process is an event that occurs internally, where it is not directly observable or interpretable. It is true that we can

3. Each component of this argument represents a complex set of issues, and so the discussions presented here are necessarily brief. But more detailed treatments may be found in Anderson (1973), Bormuth (1970 and 1969a), Bormuth, *et al.* (1970, Finn (1973), and Hively (1968).

observe the electrical effects of mental processes, but we presently have no way to interpret those effects, and most of the mental processes involved in literacy behaviors are so complex that we are unlikely to be able to do so in the immediate future.

Instead, we are forced to observe only the objects and events external to the individual to determine whether he is literate. We may observe the materials placed before him and the instructions he is given for and the questions he is asked about those materials. Then we may observe the responses he makes. So it must be recognized that what we are forced to observe in assessing literacy is not the processes that we really want to observe but merely objects and overt behaviors that we take as being signs of the presence or absence of the processes that in fact determine whether or not a person is literate. To be specific, in order to determine if a person is literate, we must have a) a theory about the nature of the mental processes that constitute literacy and b) a secondary theory that connects overt behaviors in certain situations to the various mental processes that constitute literacy.

The test task or test item is a product of this secondary theory. It functions as a set of circumstances in which a person is forced to exhibit some sort of behavior; the nature of that behavior is interpretable within the theory as evidence that the person does or does not possess the mental process being studied.

Problems with observing only overt behaviors. Quite aside from the purely scientific problems encountered in developing the theories of processes and the secondary theories of testing, there is the troublesome problem of whether it is possible to test all of the important literacy processes merely by observing overt responses to tests. For example, the critical reading skills might include a set of processes that we might label *the ability to sense ulterior motives of an author.* If a very large number of items which test these processes were devised, it would still be possible for someone to claim that many of the processes that he thinks fall under that label will remain untested by any of the items in the set and by any other items derived in the same way. This type of assertion may be used as the motive for developing new types of items. But sometimes it is used with destructive intent as the basis for the claim that testing is worthless because all testing must rely on the observation of only overt behaviors and some mental processes can *never* be observed in a person's overt behavior.

This assertion can be answered at 3 levels. First, at the pragmatic level, we can point out that the roles performed by testing are not merely peripheral to instruction but are actually essential components of it. From the point of view of the student, test items represent the only effective way he has of determining what it is that he is supposed to be learning and whether or not he is learning it. The instruction may contain many exhortations to him, telling him to strive to attain many things; but in the final analysis, the only things he *has* to learn and the only things he can *find out* if he has learned or if he needs to seek further instruction in them are just those processes required by the tests he is given. Also, from the point of view of the instructor, the only evidence he has of what he has taught or failed to teach is obtained from the tests he uses. Consequently, at the pragmatic level, the argument has little force since there remains a need to learn those processes that *can* be tested and tests are an indispensable element in that instruction.

Second, a somewhat more general argument can be built on the fact that operationalism is a fundamental prerequisite for accurately communicating scientific knowledge. A verbally expressed concept is subject to almost as many interpretations as there are people to interpret it unless that concept has been defined in terms of publicly observable events, objects, and operations. Thus, the processes underlying literacy behaviors are defined jointly by the form of the written language to be read, the form of the questions or test tasks, the relationships among the test tasks and the passage, and the conditions under which the tests are given.

At the third level, the proposition that mental processes can never be measured with overt behavior may be extended in arguments claiming that a process cannot be taught to people unless it can be tested and thus the untestable process cannot possibly be given attention at a research, development, or instructional level. A proposition that there is some important and untestable process might indeed be interesting, but it requires evidence before it can be fully believed. That evidence would probably have to be in the form of a task that would evoke an overt behavior that served to index the process in question. And finding this evidence amounts to a refutation of the original proposition that the behavior was untestable. Hence, the claim seems devoid of any substantive meaning. The principal philosophical question at issue seems to be, *Of what consequence can a mental*

process be if it has not yet been demonstrated to have any manifestations in a person's overt behavior? We cannot even make a convincing claim that such a behavior exists, without refuting that claim.

Selection of a testing mode

There seem to be just 2 major classes of test tasks used to measure literacy behaviors. The first is the *performance task*, which requires a person to read some passage and then to demonstrate a literacy behavior by performing a task that involves either concrete objects and events or pictures of objects and events. One such task might require a person to read instructions for assembling a bicycle and then have him either actually assemble a bicycle or discriminate among pictures depicting correct and incorrect methods of assembling it. The second major class of test task is the *verbal question*, which consists of an interrogative sentence requiring a response; both the question and the response are derived from the language in the passage. The person is required to read the passage and then either to write, speak, or select the response from a group of alternative responses. This type of item may range from those that ask a person to pronounce a word to those that ask him to induce and describe the moral principles that govern the behavior of the hero of a story. It should be noted that the principal distinction between the verbal question and the performance task is not whether one employs language in the test task. Both invariably do, at least in the instructions for the task. Rather, the distinction is that a verbal test question involves only language in both the question stem and the response.

Evaluation of performance tasks. The performance task superficially seems to provide the most valid type of literacy test. Perhaps the ultimate criterion of literacy would be obtained by giving a person a passage to read and then following him about through his normal life routines and observing whether the passage had the appropriate effects on his behavior. This, of course, is a preposterous proposal because of the enormous expense involved, if for no other reason. Consequently, it is necessary to employ some artificial but more convenient testing procedure and then *infer* that a correct response on this artificial task may be taken as valid evidence that the person would be found to respond correctly in his normal life routines if we were to follow him about. This can be referred to as the *pragmatic validity* of a test task.

The performance task attempts to gain its validity by simulating situations the person might encounter in his normal life routines, and it gains considerable practical usefulness because these simulations can be performed at the convenience of the tester. Still greater economy is obtained by using pictures instead of concrete objects and events. However, it should be recognized that this may reduce the item's apparent pragmatic validity, which depends on the apparent quality of the analogy between the performance task and the normal life situation. And the use of pictures may reduce this. However, whether or not a type of item is actually pragmatically valid depends solely on its experimentally demonstrated ability to predict appropriate behaviors in the person's normal life routines. Since there have been no studies attempting to demonstrate the pragmatic validity of performance items as a class, it must be said that the pragmatic validity of any performance item is apparent only, and not demonstrated.

Indeed, it would undoubtedly prove difficult, if not actually impossible, to demonstrate the pragmatic validity of the performance item as a class. To do so, we would have to define this class of items in a manner that would permit us to draw samples of items that we could be certain were unbiased representations of the total population of performance items. Then and only then could we conduct studies of their pragmatic validity, studies that would permit us to infer that the properties of the samples of items were also properties of the other items in the population. It is hard to determine even where one might start in an effort to define the population of performance items in such a manner that a random sample might be drawn from it. Possibly we might begin with a passage and identify all the situations a group of people have encountered in which they could have demonstrated their literacy with respect to that passage and then we might select from these, those situations that might be suitable for testing purposes, and finally we could study the pragmatic validity of the tasks so selected with respect to the remaining tasks. But this still leaves us wondering what it might mean to *identify all the situations a group of people encounter* and how one might go about simulating these. For the former, one would obviously have to have at least a theory of semantics that systematically related language in passages to situations; no such theory now exists. For the latter, one would require a systematic theory for relating one complex physical situation to another—another

nonexistent theory. In the final analysis, then, performance tasks have only apparent pragmatic validity, but there is very little prospect that their actual pragmatic validity can be demonstrated for the class as a whole.

The second limitation (as was mentioned above) of the performance item is the rather obvious one of expense.

The third limitation is a severe one. It is impossible to use the performance item to test the full range of literacy behaviors. A substantial amount of language is used to refer to impossible and unobservable events and objects, such as *The elf thought hard about the loss of magical powers* or *God is a disembodied power*; and some language refers to observable but extremely abstract notions such as *The search for truth is the quest for power*. It becomes difficult to imagine a way the performance task could be used to assess a person's literacy with respect to printed language of this sort. So unless it could be shown that the processes underlying responses to statements of these types were identical to the processes underlying responses to statements about real and concretely observable things, it must be recognized that performance tasks are applicable only to language that deals with concrete and observable things.

Evaluation of verbal questions. The verbal test question seems to escape most of these problems. It seems entirely possible to determine experimentally the pragmatic validity of verbal items. It is possible to develop algorisms that produce whole populations of items (Bormuth, 1970) in such a way that it is possible to either generate or select unbiased samples of items, and it is therefore possible to conduct experiments to determine the pragmatic validity of this type of item. And it was also argued in the same source that it is at least conceptually possible to develop similar definitions for *any* verbal question that is relevant to a passage.

On first analysis the verbal test question seems to involve a circularity that has an undesirable effect on certain classes of questions. The verbal task tests a person's responses to language merely by giving him a question that is also language and then observing his response, which is still more language. At no time is it necessary for the person to make a response to the objects and events referred to by that language, demonstrating that he actually understood it.

That this is not only a possible but even fairly common phenomenon can be seen from a consideration of these sentences:

1] All daxes have wobs.
2] We have daxes in our dorf.
3] Do we have wobs in our dorf?
4] What has wobs in it?
5] Who has a dorf?

Although questions 3, 4, and 5 are fairly easily answered by most speakers of English, one could hardly say that they understood sentences 1 and 2, since several of the lexical morphemes in them were in fact nonsense syllables.

However, there are many classes of verbal questions that can be defined, and *this effect* seems to be limited only to a few of those classes. Consider, for example, this sentence set:

6] The youth mounted the steed.
7] Who climbed on the horse?
8] Who mounted the steed?

It is much less likely that verbalism could occur on 7 than on 8, and each of these represents different classes of items that can be rigorously defined. Anderson (1973) has explored the evidence on this matter in some detail. Consequently, while the apparent pragmatic validity of some types of verbal questions may be questionable, those classes of items can be defined and separated from those classes of items that appear likely to be shown to have acceptable pragmatic validities.

The verbal question is fairly inexpensive to construct and use. This is not to say that the verbal questions generated by just any speaker of English would suffice for testing literacy behaviors; it requires the skills of a person highly trained in linguistics and item-writing theory to prepare acceptable items. Nor is this the claim that we already know how to write every type of item that might be employed in literacy definitions. To reach this point of development will require considerable investment in research. Rather, it is simply the claim that the verbal question will generally cost less to prepare and use than its major rival, the performance item.

The verbal question is further recommended by the fact that it is equally applicable to all language. Questions are, in fact, nothing more than transformations on the syntactic and semantic structures underlying the language in passages. Sentences 4, 5, 7, and 8 are each examples of questions derived through applying semantic and syntactic transformations to the sentence to which each, respectively,

is relevant. Number 3, on the other hand, is derived from the syntactic relationship underlying and connecting the continuous discourse represented by sentences 1 and 2. Some of the details of these question transformations have been examined elsewhere (Bormuth, 1970).

Finally, regarding verbal questions as transformations on the structures of the language in a text provides the verbal question with numerous advantages. Of greatest importance in the immediate context is the fact that these question derivation transformations enable us to give exact definitions of classes of items[4] and subsequently to use these definitions of item classes to give equally exact definitions of literacy behaviors. With respect to the performance item, it is extremely difficult to define classes of items in an exact manner because doing so requires that we possess well-developed semantic theories and theories that relate physical situations to each other—theories that are presently so poorly developed as to be almost nonexistent. As a result, it is impossible to say with objective certainty that 2 different performance items are members of the same class or of different classes. And when one cannot even say that 2 collections of items are at least formally different, there is no logical justification whatever for claiming that the mental processes tested by each population of items are in some respect homogeneous within the populations and systematically different from the processes tested by other populations. In the case of the verbal question, however, differences among classes of items can be denoted by differences in the transformational procedures by which they are derived, thereby providing at least the first logical basis for operationally defining different classes of literacy behaviors. Moreover, there is now strong evidence that the classes of questions that are generated by syntactic transformations do, in fact, test homogeneous categories of behavior (Bormuth, *et al.*, 1970).

One implication of this last statement is that the rationale and technology that underlies all educational test writing falls short of what might be considered scientifically acceptable. Thus, the benefits of treating verbal questions in a scientifically acceptable manner can be attained only after considerable effort has gone into the research necessary to lay the scientific base for the required technology.

4. The author (1970) suggested that the question transformations defined by Chomsky and others could serve as a prototype for these definitions. However, this proposal turned out to encounter several difficulties because of deficiencies in transformational grammar. Finn (1973) has since found that algorisms based on a case grammar seem to overcome most of these problems.

Tests made by traditional procedures

Having noted the verbal question as the best mode of testing literacy behaviors and having acknowledged that its value is only potential because the methods by which it has traditionally been made are not reliant on the rational procedures of a science, we should examine traditional test making procedures and some of the problems that grow out of them.

Traditional test-writing procedures. The traditional method of writing tests involves 4 steps (Bloom, *et al.*, 1956). First, the test writer lists each of the mental processes he wishes to test. These processes form a set of column headings in a table or matrix. Second, he lists all of the different types of subject matter he perceives as being taught by a passage and that he wishes to test. This list is placed in the left-hand column of the table, and each item on the list serves as a row label. This forms a table of the type illustrated in *Table 1*, where the items of content are represented by the symbols $C_1, C_2, \ldots, C_m$ and the mental processes are represented by the symbols $P_1, P_2, \ldots, P_n$. Third, he then attempts to write for each cell of his table the

Table 1 Illustration of a test writer's matrix

		Mental Processes			
		P_1	P_2	. . .	P_n
Content Items	C_1	C_1,P_1	C_1,P_2		C_1,P_n
	C_2	C_2,P_1	C_2,P_2		C_2,P_n
	.				
	.				
	C_m	C_m,P_1	C_m,P_2		C_m,P_n

type of item that permits him to test a person's knowledge of a given item of content by having him exhibit it using whatever mental process serves as the column heading. For example, suppose that P_1 stood for the mental process involved in comprehending the main idea of a

paragraph and C_2 stood for content dealing with the structure of atoms; then the item written for cell C_2 P_1 would be written in a manner that appeared to the test writer to force a person to demonstrate his ability to comprehend the main idea of a paragraph that dealt with the structure of an atom. The test writer is not provided with any definite set of operations for deriving these items. Rather, most writers on this topic (see Davis, 1964, p. 262, for example) regard the actual formulation of the item as a quasi-artistic endeavor. Finally, a jury reviews his work to see if they agree with it or if he needs to revise it.

Lack of operationalism. This conceptualization of item writing laid the basis for all modern test theory. And particularly important was the insight it gave us into the dual nature of the test item. That is, an item not only tests knowledge of some information; it also tests a person's competency to perform the processes necessary to derive that information from his instruction. However, it left a number of problems to be resolved. In one way or another, virtually all of the criticisms that can be leveled at this procedure grow out of its heavy reliance on the personal judgments and intuitions of the test writer. Or stated another way, the criticisms grow out of the absence of operationalism of the procedure—the absence of specific instructions for carrying out each step.

The test writer is told that he should test a mental process only when it is appropriate for the passage, but he is never told by what rules one decides if it is appropriate. The test writer is also told to write items that test those mental processes, but he is never told what the form of those items may be. And he is told to list the content topics he thinks the passage deals with, but he is never given any instruction on how to identify these topics or on how grossly or narrowly he should analyze these topics.

As a result of this looseness in the procedure, it seems doubtful that a test made in this way could meet the ordinary requirement of operational replicability, which is imposed on all activities laying claims to scientific status. Before an activity can be regarded as useful for making verifiable statements, we ordinarily demand that it be operationalized to the point that others working independently can perform the same operations and verify the results.

Somewhat the same demands are placed on the evaluation of programs that employ public funds and represent matters of public

policy; only in this context they are phrased as the demands for accountability, understandability, and freedom from personal bias. If 2 test writers cannot independently replicate each other's work—and it is extremely unlikely that they could—it becomes immediately apparent that the concepts of the mental processes, the subject matter topics, and the question-writing procedure mean different things to each test writer and are therefore not expressed in a form that is understandable and can be communicated. And far from being impartial, the results on such tests must be regarded as biased by whatever test writer happens to prepare the tests.

Inversion of the validity question. The traditional approach to item writing takes a peculiarly inverted approach to the question of what mental process is tested by a given item. It simply assumes that, if the test writer and his jury agree that an item measures a particular mental process, then that is, *ipso facto,* what the item tests. It does not view this as a matter that should be established by scientific procedures in which one would set out to isolate a process and study its nature, but rather as a matter to be settled only by a fiat of the test writers. Consequently, the labels on tests developed by traditional methods are highly suspect. The implicit claim that they make cannot be verified, because the test writer is permitted to use whatever label he feels is appropriate; and the lack of replicability of the work of any test writer who uses traditional test-writing methods shows that the application of these labels is highly idiosyncratic, if not actually arbitrary. Again it can be seen that tests made by these procedures have impaired value for use in scientific analyses. Similarly, it can also be seen that these tests cannot be taken as impartial evidence of the effectiveness of instructional programs or evidence that is used in making decisions that influence the lives and wealth of people since the results are likely to reflect the conscious and unconscious biases of the test maker.

At this point it may be appropriate to note that test writers themselves have long been aware of and concerned about the problems inherent in their procedures. But they have also been faced with the urgent ongoing need for tests in the schools. Consequently, they have had to do as well as they could using methods which are less than scientific until some way was found to develop better test-writing methods.

Operationalizable test-writing procedures

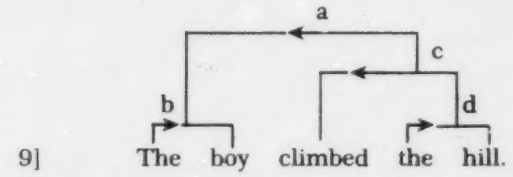
While verbal questions made by traditional procedures are of dubious value, this is not a property of the item itself; rather, it is merely a property of the way it is derived. That is, the items ordinarily produced by traditional test-writing procedures are good items in the sense that they do test some sort of behaviors that at least intuitively appear to be important behaviors. What is required, however, are item-derivation procedures that can produce populations of items in a replicable fashion. There are now 2 such procedures that may be useful for testing literacy behaviors—the cloze and the *wh-* question procedures.

Cloze procedure. The cloze procedure is a way of making tests by mechanically deleting the words in a passage of written language and replacing each with an underlined blank of a standard length. People taking the tests are expected to guess what word was taken out of each blank and write it in that space. There are a variety of ways to select the words to be deleted. One can delete every Nth word, every second noun, all adjectives, and so on. What distinguishes a cloze test from an ordinary deletion test, though, is the fact that in a cloze test the words may be selected for deletion only by a completely replicable set of rules. Using introspective concepts like *key words* is ruled out.

The advantages of such a procedure are primarily that the tests made in this fashion are completely replicable, making true validity studies possible. One can define the population of all items where predicate adjectives are deleted. And this makes it possible to draw a random sample of such items, to study their properties and then to attribute the results to that population of items. It can also be claimed that items made by these procedures do not reflect the biases of the test maker. In these respects the cloze procedure satisfies some of the most basic requirements necessary for an acceptable test of literacy.

However, cloze items have a rather serious shortcoming, because it is difficult to relate the items to the theory of language comprehension. In this theory (Osgood, 1963; Mowrer, 1954) comprehension is regarded as taking place by a series of events through which the meaning of one word or phrase is combined with or modified by the meaning of another word or phrase. And the character and order of these modifications is regarded as being controlled by the

syntax of the text. Thus, in sentence 9 the word *boy* might be thought to be modified by *the*, *hill* by *the*, *climbed* by *the hill*, and *the boy* by *climbed the hill* in that order. Ordinary verbal questions that begin



with *wh-* words like *who*, *what*, *where*, and so on can be directly related to this theory since, unlike the cloze procedure, whole words, phrases, clauses, and even sentences may be deleted. For example *What did the boy do* is derived by deleting the whole predicate *climbed the hill*. And thus the verbal question can be regarded as testing associations at each of the various points at which modifications occur in a sentence. The question *What did the boy do* tests the modification marked *a* in sentence 9, the question *What did the boy climb* tests the modification marked *c*, and modifications *b* and *d* cannot be tested, presumably because they primarily carry syntactic information rather than semantic information.

On the other hand, because only single words are deleted in the cloze procedure, only the lowest level modifications can be directly tested.[5] And some of these may primarily test structural modifications as when the word *the* is deleted. Possibly the most serious disadvantage of the cloze test is the fact that the individual test items are difficult to interpret. When we use questions, it is a fairly simple matter to relate the test item to the structure of the text, and thereby interpret what process the question tests. An inspection of cloze items, however, shows that responses to most of them depend on a variety of processes, and it is difficult to identify those processes. However, reviews of the rather extensive research literature on this type of test seem to show that what cloze tests measure is indistinguishable from what is measured by ordinary comprehension questions.

5. Some people have interpreted this fact to indicate that the cloze procedure tests only the short-range constraints in a passage. And this can be equated to testing only the simplest factual information in the passage. This interpretation, however, lacks any support in research. It is true that the short range constraints have a powerful effect on the response (McGinitie 1960), but they are by no means sufficient to fully explain cloze responses (Taylor, 1954).

Wh- questions. As just noted, the verbal questions that generally begin with *wh-* phrases such as *who, what kind of, what did,* and so on do have the desirable property that they can be related directly to the theory of comprehension. And, although in the past it was possible to derive them only by traditional test-making methods, it is now possible (Finn, 1973) to derive them by using procedures that make it possible for one test maker to independently replicate another's work and for one to give precise definitions of a number of populations of items.

This is accomplished by regarding the question as being derived from the language in a passage through a set of semantic and syntactic transformations. For example, each of the questions mentioned in the paragraphs immediately above can be derived by a set of transformations that can be crudely described as deleting one branch of a modification, replacing the deleted branch with an appropriate *wh-* phrase, and then, if it is not already there, shifting the *wh-* phrase to the front of the sentence. This description, of course, neglects many of the details of the transformations; but what is important is the fact that *it is possible to devise rules* that exactly describe how each of the various classes of *wh-* questions can be derived. And this fact makes it possible to state that *wh-* questions derived in this way seem to have all of the basic properties necessary in order to develop them into fully satisfactory tests of literacy behaviors, such as *ability to get the important points, ability to get the main ideas, ability to comprehend the structure of the knowledge or material,* and so on. The definition of such items depends on being able to assign a syntax to a passage that connects its sentences and larger segments of discourse to each other in an explicit and replicable fashion. Such an analysis seems feasible to develop at this time, and some segments of it have been developed. This syntax is then used to define various classes of questions. Bart's (1970) work strongly suggests that it is also possible to employ the syntax of Aristotelian logical algorithms to texts in order to define classes of items that test what have been known in traditional terms as the *inference skills* and many of the *critical reading skills.* This recent research provides fairly good grounds for the claim that all important tests of literacy behaviors may eventually be derived by this type of procedure.

However, one fact should be made plain. The great benefits that can be attained through this procedure of test writing cannot be

arrived at without a very considerable amount of research, and this is a type of research that the educational research community is only partially prepared to undertake. The research deals primarily with the calculi of linguistics and logic, areas of research for which only a few educational and psychological researchers have been trained. Consequently, getting research of this type under way will necessarily involve a considerable amount of interdisciplinary research and training programs.

Criterion of literate performance

The second parameter that must be specified in a definition of literacy is the level of performance a person must exhibit on a test before he may be regarded as literate. In literacy assessment we wish to perform binary classifications of people as being either literate or subliterate. And being able to do so as an either/or classification is vital, for we can then use the person's classification for making decisions that are important for him and for the society as a whole. When the individual becomes literate, he can stop using up irreplaceable fractions of his life learning literacy skills and turn those skills to more directly productive activities. And society can stop spending money for expensive instruction in literacy skills and turn its resources to other tasks. In both cases the criterion should provide a rational procedure for deciding to terminate one type of activity and to commence different activities. Unfortunately, groups of people have an annoying tendency to exhibit a continuous range of scores on tests rather than a tendency to fall into 2 well-separated clusters of scores. Consequently, there is no natural or immediately obvious way to make the binary decisions required.

Problem

Thus it can be seen that what we are really dealing with at this point is the classic problem of *How good is good enough?*—where goodness is measured along a continuous scale having no natural boundaries that would facilitate metrically clean and logically neat binary decisions about when a person's performance on literacy tasks is good enough to warrant his being labeled as literate. Or, stated operationally, the problem is to assign social values to test scores and then to identify the score that has the greatest value to a person. This

problem can be solved when it is properly conceptualized. But first it may be instructive to look at previous approaches to its solution.

Earlier criterion scores. At various times a number of criterion scores have been advocated. Each has been the subject of some scepticism for both technical and philosophical reasons. Let us begin by noting 3 of the better-known criterion scores. One of the earliest and most widely used criterion scores was proposed by E. L. Thorndike (1917). He recommended that a student should be able to answer at least 75 per cent of the questions on a test made from a sample of an instructional material, if that material were to be considered suitable for use in that student's instruction. Instructional programmers sometimes employ the so-called 90-90 criterion wherein they attempt to revise and improve their materials until 90 per cent of their students can answer 90 per cent of the questions given them at the termination of instruction. Finally, some have interpreted (probably erroneously) the writings of Bloom (1968) and Mayo (1970) as advocating that a student's instruction in a body of content be continued until he can demonstrate perfect performance on a test of that content.

Nearly all proposals of criterion scores, including these 3, have assumed that tests made by the traditional procedures would be used to measure attainment of the criterion level of performance. Lorge (1948) pointed out that the absolute magnitude of scores on tests of this type are usually subject to biases introduced by the ideosyncracies of the people who wrote the tests. Test items testing different content and processes often differ widely and systematically in difficulty. In addition, even slight variations in the phrasing of a test item can sometimes lead to wide variations in the difficulty of the item. Thus, a test writer can have a great deal of influence on the difficulty of the test. In traditional test-writing methodology, the test writer is only partially constrained with respect to the content and processes that he may test and almost completely at liberty to phrase his items to suit his personal preferences. Hence, 2 different test writers might be expected to produce tests of quite different difficulties to test exactly the same instruction. Thus, Lorge reasoned that these criterion scores do not represent a standard level of competence.

Few advocates of a criterion score have advanced a rationale to justify their preference for the scores they chose. Bloom (1968) appears to have been the exception. He pointed out that students required to reach a high (but not necessarily perfect) level of per-

formance on each unit of content tend to exhibit similar levels of final achievement, to overcome initial deficiencies in aptitude, and to master succeeding units of content in less time. Bloom did not specify a particular level of performance as a criterion. He simply advocated using a high level of mastery as a criterion. However, many have interpreted his concept of mastery to mean perfect performance. And a critique of this misinterpretation is highly relevant to the problem we are considering here.

There are at least 3 reasons why perfect performance is unlikely to be considered the most desirable level of performance. First, using such a criterion is likely to drive some of the costs of instruction to preposterous levels. Almost every learning study shows that learning increases rapidly on the first few repetitions of the material and then flattens almost to the horizontal well before perfect performance is reached. Hence, attempting to reach perfect performance is likely to be a time-consuming and expensive undertaking. Second, since efforts to reach this criterion are likely to involve much repetition and drill, we could anticipate adverse effects on the students' attitude toward the content of instruction. That is, students find much repetition boring and unpleasant, and they could transfer those attitudes to the content and thereafter try to avoid its further study and use. Third, attempting to reach perfect performance on a unit of content implies that all of the items of content in that unit are essential to learn. This may be true of a few isolated units, such as that dealing with the multiplication tables, but most units of content deal with collections of content items that differ greatly in their utility to the individual.

Reformulation as a rational problem. This critique of the criterion of perfect performance now puts us in a position to reformulate the problem of identifying criterion scores. The authors of most criterion scores stated a preference for a particular score but failed to support their preference with reason and evidence. This seemed to reduce the problem to simply making an arbitrary choice of a score. However, we have just seen a case in which plausible arguments were offered in favor of setting a high criterion score and equally plausible counter-arguments were advanced against setting the score too high. Hence, the problem will clearly submit to rational analysis and formulation.

Let us start with the proposition that neither literacy behaviors nor any others are taught for the pure hedonistic pleasure of learning. Rather, they are placed in curricula because they are valued for the tangible and intangible things that those behaviors enable individuals and society to attain. But, even then their placement in curricula is decided only after these positive benefits have been weighed against the negative benefits, that is the costs associated with learning the behaviors. Second, an individual can acquire varying numbers of literacy skills, and this fact can be accurately indexed by his scores on an appropriately made achievement test. Third, each level of performance produces effects on each of the various benefits, both positive and negative, that he is likely to accrue. Finally, the problem of identifying a criterion score, then consists in finding the level of performance at which the over-all benefits are the greatest.

An ideal solution

This problem can be approached at 2 levels—either through an attempt to describe how one might establish a performance criterion right now, using practical procedures presently available to us, or through an analysis of the operations required to attain what might be thought of as the "ideal" solution. Both courses were elected because a description of the ideal solution provides a framework within which to evaluate the adequacy of any practical solutions proposed. A practical solution will subsequently be described and developed in some detail.

In the ideal solution to the problem of establishing a performance criterion, one should be able to attach a cost and a benefit weight to each literacy behavior separately, select those behaviors whose learning seems likely to produce a net positive benefit for the learner, and then determine what level of performance on a test of these behaviors is associated with the greatest expected benefit to the learner.

It is useful, of course, to consider the value of whole classes of literacy behaviors as units, because one of the major attributes by which we categorize literacy behaviors is the subjective value we assign to the function that those behaviors ordinarily perform for us. Specifically, classes of literacy behaviors have not traditionally been defined *solely* in terms of their psychological attributes but also in terms of their functions and how we value those functions. For ex-

ample, being able to identify the plot of a story seems, at least to me, to be no less a cognitive behavior than being able to identify the outline or rhetorical patterning of an expository essay. And the 2 classes of behavior seem likely to share many elements in common. However, these 2 types of behaviors do function quite differently, and their respective functions are valued quite differently. Identifying the plot of a story functions as a skill intended to enhance aesthetic appreciation of literary materials, while identifying the rhetorical pattern of an essay functions as a skill in more utilitarian tasks.

But not all of the behaviors within a category are equally useful or equally expensive to teach. For example, some phonics rules apply to many different words that occur frequently in the language, while other rules seldom apply; and thereby these rules can be said to differ in value. Similarly, the syllabication rules, which could be very useful in phonics behaviors, seemingly cannot be taught as fully effective reading skills until students have learned to discriminate between English words of Germanic and Romance language origins, an operation that seems likely to be so costly that no modern scholars are seriously advocating it. Doing so involves too great an expense in view of the rather limited benefits to a reader.[6] Thus, when the individual behaviors of which phonics is composed are analyzed, each may be assigned a different set of benefit and cost values. Consequently, before we can decide on a criterion level of performance for a given category of behaviors, we must give careful consideration to selecting the individual behaviors to appear in that set and the values to be assigned to each.

It would be both ideal and very convenient if we could then simply sum these cost and benefit values across behaviors to arrive at performance criterion scores. And, if we had complete knowledge of the nature of all literacy behaviors and of their relationships to each other, we could undoubtedly perform such an operation with a fair degree of confidence in the results. Unfortunately, at the present time we cannot, for example, even identify all of the literacy behaviors. Nor do we know how the word pronunciation behaviors relate to each other. It is quite commonly observed that a word is more easily pronounced when it appears in context than when it appears out of con-

6. Statements about print-to-sound phonics (*i.e.*, reading phonics) made here should not be confused with or generalized to apply to statements about sound-to-print phonics (*i.e.*, spelling phonics).

text and consequently that context guessing skills interact with other word recognition skills, with the result that including one type of skill in the instruction probably influences the costs and benefits arising from including others. Hence, it is impossible at this stage of our knowledge to arrive at a performance criterion by simple summing operations, at least not without doing a considerable amount of research.

A second practical problem stems from the fact that the literacy behaviors have not been analyzed to the point where we can separately identify all of the important processes involved in literacy. Indeed, it is not even clear what is meant by *all of the different processes*. Two processes differ when they contain either different components or the same components differently related to each other. And it is at least conceivable that we could analyze a process until we had identified the activities of individual neurons and the sequences of those neuronal activities. If we did so, we potentially would have a very large number of *different* processes. But obviously an analysis thus detailed would be extraordinarily awkward for use in instruction, since it is neither necessary to use so fine an analysis in instruction, nor practically possible to operationalize the instructional and testing procedures of each of the different processes identified at this level of analysis. Consequently, deciding what literacy behaviors should be taught and tested depends on considerations of how far it is necessary and desirable to analyze the processes underlying these behaviors in order to obtain instruction with the desired level of effectiveness. Thus, it is difficult to see how we could establish a performance criterion on each individual literacy behavior.

A third problem that will have to be solved is how to demonstrate the pragmatic validity of the items in a literacy test that is intended to exhaustively test a category of processes. A person is literate with respect to a particular real-world reading task if he can perform it competently. Yet we are proposing not to observe him perform that task but rather to analyze the performances required by all such tasks into the abstract processes underlying them, to operationalize processes using test items that differ in many respects from the real-world reading tasks, and then to infer that some level of performance on our test of the abstracted processes permits us to claim that he would perform competently on the real-world tasks. Demonstrating the validity of these inferences amounts to a demonstration of the

validities of the theories about the processes underlying the literacy behaviors. Hence, when this argument is pursued, it can be seen that it is impossible to achieve the ideal performance criterion until a very great amount is known about the literacy processes, a goal which probably will not be achieved without a fairly large amount of research.

Procedures presently practical

All this is not to say, however, that it is impossible to greatly improve present test-writing procedures. We are in fact, developing 2 testing procedures that remedy many of the weaknesses of tests constructed by traditional procedures, and we can employ these procedures a) to establish rational performance criteria for at least 2 of the major categories of literacy behaviors and b) subsequently to incorporate these criteria into literacy assessment procedures that are substantially better than those in current use. This section will describe the first step in this procedure, and subsequent sections will describe the remaining steps.

Basic design of the procedure. The approach suggested here begins by accepting the unpleasant reality that we have very incomplete theories about the processes underlying literacy behaviors and, therefore, that we cannot at the present time have operational procedures for testing those processes in such a manner that each process is individually identifiable. Instead, it asserts 1) that we do possess operational testing procedures that seem to test most of the word recognition and literal comprehension processes involved in literacy behaviors, even though these testing procedures do not permit us to isolate each individual process; 2) that these test-making procedures are adequately operational for establishing rational performance criteria because they do not permit test writers to bias the tests; and 3) that these rational performance criteria can be incorporated into a literacy assessment design that will produce results that, though admittedly short of ideal, are more believable than any produced by the traditional methods of assessing literacy.

In the approach that seems practical at this time, the criterion functions somewhat differently than in the approach already described. Instead of attempting to abstract underlying processes and to attach a value to each one separately, the approach suggested here proposes to use testing procedures that seemingly test a variety of

reading processes. The reading tasks are selected to have direct practical significance, and then research is conducted to identify the most desirable score for people to attain on these tests. It will be noted that this approach requires, at the very least, both operational test-making procedures that can be applied to any passage and that do not permit the idiosyncrasies of test writers to bias the results, and a theory for deciding what is the most desirable score (the performance criterion score) on any test made by this procedure. Thus, in this approach one simply attempts to develop procedures that determine whether a person is literate with respect to a *specific, important, real world reading task*. When this has been done, we can then employ this operation to determine any person's literacy with respect to any set of real-world reading tasks.

Criterion identification model. Now, let us consider the problem of how one decides which score on a test is the most desirable score. This topic is necessarily abstract. However, it can be made more comprehensible by illustrating it with data from a series of studies that I am currently conducting. The objectives of this study are 1) to develop a rational model for identifying criterion scores, and 2) to employ the model to identify criterion scores for use in interpreting scores on cloze tests that are employed for evaluating the comprehensibility of instructional materials.

The reasoning behind this model is fairly simple. People read because doing so produces several effects on them, and they value those effects. Hence, if one of the scores on a test made from a passage is any more desirable than any of the other scores, that score is more desirable because it is normally associated with a greater total value arising out of these effects. To be a bit more specific, the value (V) of a given score (i) on the criterion test (C), in this case a cloze test, is given by

$$V(C_i) = (w_1 \cdot E_1) + (w_2 \cdot E_2) + \ldots + (w_x \cdot E_x)$$

where w_1 stands for the value we place on effect number one, and E_1 stands for the amount of effect of type one that we normally expect to find associated with cloze score i. The value derived from each individual effect is obtained in exactly the same way that we calculate the value in any other accounting problem—we multiply the number of units of E_1 that we obtain by the average value (w_1) of each of those units. Thus, this model simply claims that the value of a given test

score is the sum of the values of each of the individual effects that we normally expect to find associated with that score. The criterion score for the test is the score having the greatest summed value (V). Note that this model is quite general and can be applied to any test, regardless of how the test is made and regardless of what that test is measuring.

Seven steps are required in order to apply this model. First, a criterion test (C) must be selected. This illustration employed the cloze readability test. Second, the associated effects (E_1, E_2, etc.) must be identified and tests must be developed to measure them. These include both negatively and positively valued effects. Third, the criterion test and the measures of the other effects are administered in a manner that permits each effect for a person to be associated with his score on the criterion test. Fourth, the average amount of each effect is calculated for each score on the cloze test. Fifth, the people who are best able to estimate the relative values of each of the effects are identified, and we obtain the average of the values that they assign each effect. Sixth, each effect is multiplied by its value. Seventh, for each cloze score, the values of the individual effects are summed. Eighth, the cloze score having the greatest value is found and designated as the criterion score.

Application of the model. In the study[7] that will serve to illustrate the application of this model, 4 effects were identified: Information Gain, Rate of Reading, Conceptual Difficulty, and Willingness-to-Study. Information Gain was measured by testing a person's knowledge of a passage both before and after he had read it and then calculating a residual gain score for him. The tests used to do this were made by a fairly operational procedure that involved using the sentence and intersentence syntax of a passage to transform the text into questions and then drawing a stratified random sample of the questions that resulted. Rate of Reading was measured by having a student read a passage, noting how much time it took him to complete the task, and then calculating the number of words read per minute. Conceptual difficulty was measured by having the student read a passage and then rate it on a 7-point scale in which the extremes were labeled *much too easy* and *much too difficult*. The scale was then

7. This study employed a very complex design that involved the use of passages of matched difficulty, counterbalanced rotations of materials, and the like. If the reader is interested, he can find the details in a preliminary technical report (Bormuth, 1971).

folded so that the scores ranged from 1 to 4, from *much too easy-difficult* to *about right*. Willingness-to-Study was measured by having the student read and then rate a passage on a 7-point scale that ranged from *like very much* to *dislike very much*.[8]

The value of each effect was determined by having the teachers of these students rate the behaviors. The teachers were given a description of each of the tests and asked to rate the relative values of the behaviors on a 10-point scale. The average rating given each behavior was then used in the model.

A sample of the results is shown in Figure 1.[9] This figure might have been built up in this way. Consider a cloze score of 50 per cent, for example. The average Information Gain scores of students who obtained 50 per cent was calculated[10] and then multiplied by the weight the teachers had given it. This yielded a value for that effect of about 11. (This number has only a relative meaning.) This point could then be plotted and it would fall on or very near the lowest curve on the graph, the curve labeled *Information Gain*. This same process could next be repeated using the mean rate score and the weight assigned to that variable. This amount of value could thereupon be added by measuring up from the point on the Information Gain curve. This would produce a point that falls on or very near the second curve labeled *Rate*. The process could then be repeated for each of the remaining 2 measures to obtain points that fall on or near the 2 uppermost curves. *The point on the top line of this plot represents the total value a reader could normally be expected to receive from a passage if he were able to obtain a cloze score of 50 per cent on that passage.* By repeating this process for each of the other cloze scores and connecting the points obtained for each effect, we would obtain the curves shown in this figure. Since the top curve in this graph represents the total value associated with each cloze score, the point at which that

8. Explicit instructions accompanied these rating scales. This study used tests made from 32 passages representing 8 levels of difficulty and administered those tests to 1,600 students, who were drawn in equal numbers from grades 3 through 12.

9. It should be cautioned that this figure grossly over simplifies the results. The curves for each of the effects differed with the grade levels of the students, and the weights assigned to each of the effects and the curves for the rating scales differed when they were analyzed according to the type of reading assignments in which the given materials was supposed to be used. Therefore, the criterion score that can be identified with this figure will vary somewhat from the criterion scores that are actually appropriate for students at various grade levels who are asked to perform various kinds of reading tasks.

10. The scores for all of the effects measured were transformed into standard scores to remove arbitrary scale effects.

curve reaches its highest point (at a cloze score of approximately 45 per cent) represents the criterion score.

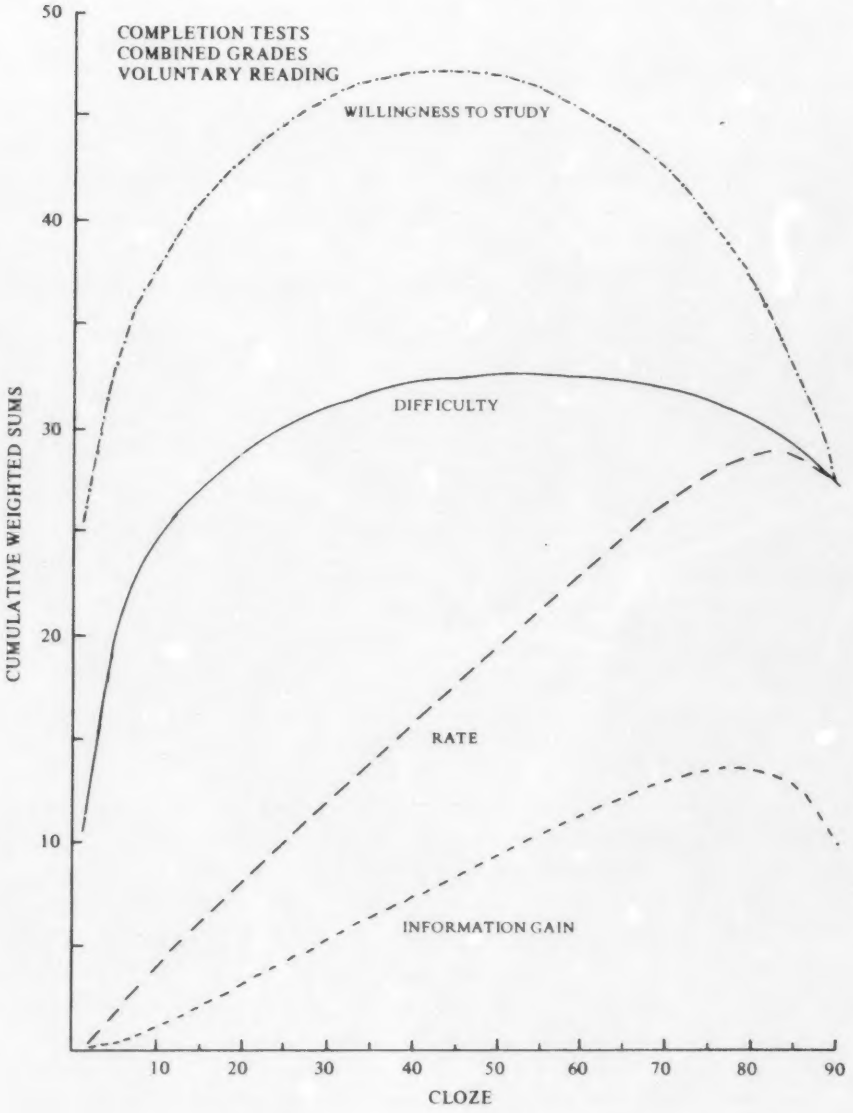


Figure 1

Formal model. In a general discussion of this sort it was undesirable to go into much detail on matters such as the rationale of the model, the instrumentation of the tests, the design of the studies or the treatment of the data. These will be made available in future publications. However, it may be desirable to present the exact form of the model for the reader to evaluate. It is given by the expression

$$V(C_i/B_1, B_2, \ldots, B_b) = (w_1 \cdot f_1(C_i)) + (w_2 \cdot f_2(C_i)) + \ldots + (w_x \cdot f_x(C_i))$$

In this expression

V = total value normally associated with a given score on the criterion test;

C_i = the score i on the criterion test C, in this case the cloze test;

B_b = the boundary conditions number b. Boundaries must include those factors that define the domain within which criterion score is applicable, factors such as the age of the student, the subject matter of the material, and the purpose for which the material is read;

$V(C_i/B_1, B_2, \ldots, B_b)$ = the total value, V, normally found to be associated with the score i on the criterion test C, within the boundary conditions $B_1, B_2, \ldots, B_b$;

w_x = the relative weight normally assigned to effect x, effects such as the person's average expected income, the proportion of news editorials that the person is likely to be able to read as a desirable level, etc.;

$f_x(C_i)$ = the amount of effect x associated with the score i on the criterion test as given by the regression of that effect on the criterion test. This measure is expressed in some standard form to remove arbitrary scale effects.

Four characteristics of this model should be noted before we leave the subject. First, almost any kind of test can be used as the criterion test, including the typical standardized achievement and aptitude tests. The metric and content of those tests is fairly arbitrary in the sense that the items contained in them are selected primarily to produce a particular set of metrical characteristics for the test and only secondarily to test a "representative" sample of some body of content. However, only operationally defined tests may be used to measure what we have been calling the effects in the model. A criterion score is normally generalized to a population of instructional stimuli and this cannot be done legitimately unless we can be assured

that fairly comparable measures were applied to the sample of stimuli originally used to identify the criterion score. Second, the model is not biased by including irrelevant effects. An effect may be irrelevant either because it is unrelated to the criterion measure or because people place a zero value on the effect. In either case, the criterion score identified will be unaffected by including an irrelevant effect. Third, the validity of a criterion score identified with this model diminishes, depending on the proportion of the relevant effects included in the model, on how highly the excluded effects correlate with the criterion, and on how much value people place on effects excluded. Finally, it should be noted that the model in its present form does not attempt to reconcile the trade-off between future and immediate benefits. But since the accounting procedures for doing so are well known and since those procedures were not particularly relevant at this point in the discussion, no effort was made to incorporate them into the model. However, it should be noted that they are relevant for other uses of this model.

Identification of corpuses of tasks and a corpus criterion

The third parameter that must be described in a definition of literacy is the kind of language with regard to which people should be literate. The goal of a literacy program would be hopelessly vague unless the definition on which it was based contained such a description. Language is used for many areas of discourse and for many different purposes within those areas. The language employed in each area differs materially in vocabulary structures, in sentence structures, and in discourse structures; and each of those different structures presumably requires a different literacy process or level of skill to cope with it. Thus, if a definition failed to specify the corpus or population of reading tasks it dealt with, it would implicitly commit the program to deal with all possible corpuses of discourse—a task of overwhelming magnitude.

In view of the facts that some omissions must be made for practical reasons and that the omission of literacy skills from a literacy program can have important social consequences, the position taken here is that a literacy definition is both dangerously vague and logically incomplete unless it specifies the types of language with

respect to which people are expected to demonstrate literacy. There are 2 aspects of language population selection that must be discussed. The first is a discussion of what criteria should be used to select language populations in order to reflect accurately the values of society and the individual. The second is a discussion of a procedure for establishing a criterion score for determining if a person is literate with respect to a population of written language.

Criteria for selecting corpuses

The procedure for selecting corpuses of language and the rationale for that procedure must be made as explicit as possible. Literacy skills may be employed in a number of important political, social, cultural, and economic activities in our society; and having or not having those skills has direct implications for a person's rights, responsibilities, and opportunities to participate in those activities. For example, it is not uncommon to hear that persons and even whole groups of people vote for candidates who are actively working against the best interests of those people or of people who failed to receive job promotions because they could not acquire the information necessary to carry out their new duties. Since situations of this kind are traceable, at least in part, to those people's failures to acquire certain literacy skills, it must be recognized that both society and the individual have the right to know exactly what literacy skills are selected or excluded from instruction and the reasoning by which these decisions were made. Or, stated another way, these decisions cannot, as they have been up to the present, be left obscure by allowing them to be treated as the creative acts of individual teachers or as the unexplained technical decisions made by publishers of instructional materials.

This matter has not previously received the careful analysis it deserves, and the present discussion will merely pose the problem and demonstrate the need for its further analysis. At least 5 criteria seem relevant to making decisions about whether to include or exclude a corpus of reading tasks: 1) monetary cost, 2) economy of time, 3) value-achieving utility, 4) commonness, and 5) frequency.

The first criterion that must be considered is how much money is available for teaching literacy skills and how much it will cost to teach each person enough literacy skills so that he is literate with respect to a given corpus of reading tasks. While the monetary

cost involved in literacy programs does not in any sense represent our highest value in these matters, it nevertheless, sets rigid bounds within which our other values may be achieved. At the present time, it is not clear how one might go about obtaining an accurate estimate of the cost of instructing people in literacy with respect to a particular corpus of tasks. The cost would depend not only on the nature of the tasks, but also on what other tasks were being taught and the sequence in which they were taught, because there would undoubtedly be major transfer effects among different corpuses of tasks. However, it is clear that this criterion warrants very careful attention because the resources expended developing literacy for one class of tasks necessarily preempts the use of those resources to develop literacy on other tasks.

Second, we must consider the amount of instruction time that must be devoted to achieve literacy on a corpus of tasks. Perhaps because instruction usually produces large long-range benefits for an individual and because education in the past has been under-supported, we tend to think of instruction as a general good that we can never get enough of. But this notion must be carefully examined, for instruction inevitably consumes a substantial part of a human being's most valuable and irreplaceable resource, his life. With a major and growing proportion of our population now enrolled in formal schooling for as much as a quarter to a third of the average human life span, educators cannot for much longer continue to treat the time spent in instruction as if it were a valued but essentially inexhaustible commodity. Hence, we cannot use this resource indiscriminately for teaching literacy skills, but must ask whether the benefits derived from literacy instruction on a class of tasks really warrant the expenditure of time when we consider the other literacy skills the student could have been acquiring and, for that matter, the other educational and noneducational activities he could have been engaged in. Making the time estimates required for applying this criterion promises to present problems that are similar to and at least as complex as those involved in making monetary cost estimates.

The third criterion is the value-achieving utility of acquiring literacy skills on a set of reading tasks. Each corpus of reading tasks can be employed to achieve some sort of social, political, cultural, or economic values for a person. These values should determine assigned weights that correspond to some consensus of their relative impor-

tance, and each class of reading tasks should then be evaluated and ranked in terms of these values.

The fourth criterion is the degree to which all people must deal with the class of reading tasks—the commonness of the task. A considerable degree of economy may be achievable by separating those tasks on which everyone should be literate from those tasks associated with special occupations and hobbies. Only those tasks that are commonly needed by all should be included in the definition used with a basic literacy program to be conducted for everyone. The specialized tasks may then be included in definitions of special literacy programs designed for those who seek specialized training.

The fifth criterion, the frequency with which a type of task is encountered, appears to have a dubious value for the selection of corpuses of reading or literacy tasks. According to this criterion, one would assign each class of tasks a value corresponding to the frequency with which a person must deal with the tasks of that kind and then select just those classes of tasks that occur most frequently. For example, by this criterion one would select the tasks involved in reading newspaper stories about foreign affairs while perhaps excluding the reading of diplomatic position papers dealing with similar events. The fallacy inherent in this criterion is that some tasks, such as reading the fine print in a sales contract or a sign saying *high voltage*, may occur so rarely that they would be excluded by this criterion but have consequences so critical that they could not be ignored. If this criterion is employed at all, it should be done only with caution.

Criterion of corpus literacy

Deciding that a person should acquire literacy on some corpus of printed language presents us with the familiar question of what level of performance we are willing to consider a satisfactory level of performance, and with the problem of measuring that performance.

Instrumentation. It was pointed out earlier that a criterion score can be identified using almost any kind of test as the criterion measure and that even the typical standardized achievement and aptitude tests, the so-called norm-referenced tests, can be used for this purpose. This statement should be qualified somewhat at this point. Great care is often taken in constructing these tests to develop a pool of test items that actually test a domain of content and to represent that domain of content adequately. However, the chief function

of these tests is to discriminate reliably among the people to be tested. Consequently, when the test is composed, the items selected are usually selected on the basis of statistical criteria. Miller (in press) has conducted a study using data from the Venezuelan National Assessment of Mathematics skills in which he found that applying these criteria seriously reduced the extent to which the test items could be said to represent the content of the instruction. Items representing some major blocks of content were largely and even wholly eliminated from tests while the items representing other blocks of content were vastly over-represented in tests. The proponents of the current norm referenced tests often counter by arguing that their tests, nevertheless, yield such high correlations with tests that do represent a domain of content that the results are indistinguishable. This claim seems fairly likely to be borne out. However, we still cannot completely discount the counter claim that, regardless of the size of these correlations, the norm referenced test may misrepresent substantial blocks of relevant content and, therefore, does not actually represent the content domain in the way that a criterion test is normally expected to represent it.

Scaling a standard criterion test. There are 2 major operations involved in this procedure. The objective of the first is to estimate the distribution of the cloze difficulties of the tasks in the corpus of tasks being considered. The object of the second is to scale a standard criterion test so that scores on it can be used to estimate the proportion of passages on which a person is literate. The first operation consists in selecting a fairly large random sample of passages from the corpus of reading tasks, making a cloze readability test over each, administering these tests to subjects, calculating the difficulty of each passage, arraying these difficulties in a distribution, and then assigning a percentile score to each passage's score along this distribution. Each person tested is also required to take a test selected to serve as a standard criterion test.

The second operation involves several steps that result in a 2-column table. The first column contains the raw scores on the standard criterion test and the second column contains a percentile score corresponding to each raw score. These percentile scores show the proportion of cloze tests on which the average person receiving the corresponding raw score was able to score at least as high as the criterion level of literate performance. The validity of the operation rests on several unreported studies in which the author has consistently

found that, when scores on several cloze and other operationally defined tests are regressed on the scores from either another cloze test or a standardized test of reading achievement, the slopes of these regressions are essentially parallel. Ceiling and floor effects often distort the distributions of scores, but logit transformations render the slopes parallel.

The first step of this scaling operation is to regress each set of test scores on the standard test, performing a logit transformation whenever ceiling or floor effects are present. The second step consists of calculating the equation for each of these regressions. In the third step the criterion score selected in the manner described in the previous section is substituted in each equation and the equations solved to determine, for each passage, the raw score on the standard test that corresponds to the criterion score on the test over the passage drawn from the population of tasks. The last step is to take the percentile score assigned to that passage and assign it to the raw score calculated from the regression equation on that passage. For purposes of clarity, the various details of these calculations have been omitted.

Thus, when this standard test is administered to some new subject, we can use it to estimate his corpus literacy level—the percentage of tests on which this person could achieve a criterion level of performance if new passages drawn from the same corpus of tasks are tested.

Selecting a corpus criterion score. Although it is undoubtedly informative and possibly even useful to know the proportion of passages on which a person is literate, we must again raise the familiar question: *on what proportion of the passages should he be literate? Or how good is good enough?* And the answer again seems to rest on the development of a decision theory that permits us to consider simultaneously the relevant negative and positive benefits associated with each level of literacy and to obtain a criterion score that maximizes the values we wish to derive from literacy instruction. This will be referred to here as the *corpus criterion score.*

This procedure should utilize the criterion identification model and proceed much as we did in the previous illustration. First, we would identify the negative and positive benefits associated with being able to read various proportions of the corpus at a desirable level. These might include estimates of a person's expected income level and other measures of his occupational success, estimates of the

costs associated with raising people to each level of literacy, estimates of the degree to which people can and do participate effectively in political-civil affairs in cultural activities. These measures would be assigned relative values.[11] And finally, the data would be entered into the model. The model would thereupon identify, for some hypothetical average person, the highest level of performance on that corpus having a positive value.

The reader is again reminded that the notion of a corpus criterion score cannot be fully acceptable unless certain characteristics of individuals and populations of people are taken into account. The treatment given here was intended only to develop the concept of corpus criterion scores and to illustrate a rational procedure by which they can be identified using whatever test procedure might be suitable and available.

Identification of characteristics of individuals in the population

The fourth parameter of a literacy definition consists of a set of characteristics of the people who are the subjects of the literacy program. It is almost a cliché to point out that there are individual differences among people, differences in their native endowment, environmentally acquired assets, and motivations. Hence, there is every reason to believe that their instructional needs will differ, not just in how much instruction should be administered, but also in which literacy skills they should learn and how many of those skills should be mastered. While the preceding sections have not entirely ignored the characteristics of the individual, neither have they examined them systematically. The present section will examine why individual characteristics must be represented in a model, identify the major variables, and then present the outline of the model as far as it will be developed by this investigation.

Inclusion of individual characteristics

The objective of a criterion model is to help us make decisions about the instruction of people—decisions that will help people to realize their aspirations while simultaneously conserving their re-

11. Presumably these values would be appropriately adjusted by the usual accounting procedures so that they would accurately reflect the trade-off between immediate and deferred benefits.

sources. It is doubtful that a model that omitted individual characteristics could reach this objective satisfactorily. To omit them would have the effect of forcing us to apply the same criterion to everyone, or worse, to allow only a select few to acquire literacy skills. And these skills would be the ones that were found to be appropriate for the *average* person. This would force us into an enormous waste of personal and social resources and to achieve results that had little correspondence to the aspirations of individuals and society. Consider the obvious case of the severely mentally retarded child who cannot possibly aspire to accomplishing much more than bare self-sufficiency in the simplest kinds of occupations. Almost any criterion that is appropriate for the average person in a broad segment of the population would certainly include many skills that the mentally retarded person would have little occasion to use. Moreover, because he learns very slowly, his instruction would be very prolonged, expensive to society, and expensive to him in terms of the proportion of his life that he would have to devote to the learning task. Conversely, consider the mentally gifted person who can aspire to occupations of great complexity and of great benefit to himself and society. He could attain this criterion rapidly and at little cost to anyone, but his instruction would be terminated well before he could realize his aspirations. Hence, it should be clear that the cause of neither justice nor efficiency is served by applying the same criterion to everyone.

Identification of relevant characteristics

Three factors seem particularly relevant in identifying a literacy criterion for an individual—his native capacity to learn, his environmentally acquired capacity to learn, and his motivations. We will begin by discussing the need to distinguish between the first 2 of these factors and then proceed to discuss each factor separately.

Distinction between native and acquired capacity. In discussions of instruction it has been customary to lump together native and acquired capacity to learn under the single labels *intelligence* or *aptitude*. In part, this confounds the results because it has been difficult to separate the 2 in practical measurement operations. In tests we use problems of various sorts to measure aptitude for learning and responses to those tasks rest on both native and acquired abilities. However, it now seems essential to distinguish between the

2 concepts both conceptually and in practical measurement operations. Our society is now making unmistakable demands that education give greater priority to helping each individual develop his potential regardless of his social-environmental background. Thus, educators are being asked to distinguish between biologically and environmentally acquired capacity to learn, to adapt instructional content and methods correspondingly, and then to allocate educational resources in such a manner that a person's social environmental circumstances no longer serve as a major determinant of the educational level to which he can aspire. Since the criterion model proports to identify for each individual the level of literacy to which it is most desirable that he aspire, it follows that this distinction should also appear in the model.

It may be objected that the distinction is futile because we currently have no way to assess the 2 concepts separately. This is true, but we might be able to solve the problem in at least a modestly satisfactory manner. We know many of the social-environmental factors that correlate with scores on intelligence and aptitude tests—the educational attainment of the parents, parental income, and so on. And we have estimates of the degrees to which these factors are themselves heritable. Consequently, we could weight these factors in a person's environment with a degree of confidence and partial them out of his test scores, thereby obtaining separate estimates of the biological and environmental components of his learning capacity.[12]

Native capacity for learning. The primary reason for including capacity in the model is that it is a major determinant of what it costs the person and society for him to master a given body of content. Carroll (1963) has shown that aptitude can be operationalized in either of 2 ways: by the amount that a person is likely to learn with fixed amount of instruction or by measuring the amount of time required for that person to reach a criterion level of performance in learning some body of content. From the point of view of this model, we are most interested the conceptualization of capacity to learn as *time to reach a criterion*. Time spent in instruction can be translated directly into costs. The instruction costs the student an irreplaceable

12. The reader should be alert to the fact that much controversy presently swirls about the genetic heritability of mental abilities, controversy that may make it socially unacceptable to represent the distinction between native and acquired ability in a model. However, if abilities are genetically determined to any important degree, this makes it clear that that "fact" can be ignored only at the cost of visiting considerable injustice upon those individuals who were poorly endowed by their parents.

fraction of his life, a fraction of his life that has value to him to the extent that he could be using it to produce other things that would also be satisfying to him. Similarly, the time spent in instruction is a major source of the costs of education to society. Thus, we cannot very well identify a criterion without considering the individual's native capacity to learn.

Acquired capacity. A person's acquired capacity to learn affects costs in much the same manner as biological capacity. But acquired capacity must be weighted differently in the model, since deficits of this kind can be overcome through instruction and since society is willing to allocate a considerable portion of its resources to overcoming them.

Motivations. People differ in the kinds of occupational choices they make and in the kinds of cultural and political activities that they engage in. Each of these pursuits involve different kinds of reading tasks and different amounts of reading. Consequently, if this factor is not taken into account, much could be wasted by teaching people skills that they neither wanted nor would ever use or by failing to teach them skills that they needed. Let us refer to this concept of motivation as the individual's *intentions* and distinguish it from another sense in which the term is used.

The term motivation is also used to refer to the extent to which a person perseveres in attending to a learning task. This is an important consideration in the model since it also helps to determine the costs of the instruction. If a person perseveres in attending to the instruction, the costs will be lower than if he does not.

These 2 concepts of motivation should probably be represented quite differently in the model. Intentions should probably be represented in the boundary conditions for 2 reasons. First, a person's choice of pursuits is not a continuous variable. Rather, it is simply a person's choice whether or not to enter each of a number of different pursuits that have no obvious dimensional continuity. Second, this variable undoubtedly interacts with the weights assigned to the effects that are associated with various levels of mastery of literacy skills. That is, the weight that we would assign to Information Gain, for example, depends to some extent on the reason that motivated us to read in the first place. On the other hand, a person's perseverance is a continuous variable and can be treated much as any other variable.

Form of the model

Thus, when we formalize the model as it now stands, we obtain the expression

$$V(C_i^s/1_j) = f(N_w, A_w, P_w, E_w^d),$$

which is to say that the value (V) that a person is likely to accrue as a result of achieving a given level of performance (i) on a criterion test (C) which measures a given set of comprehension skills (s) and given that he intends (I) to elect a particular pursuit (j) is some weighted (w) function of each native capacity to learn (N), his acquired capacity to learn (A), his ability to persevere on the learning task (P), and the appropriate weighted and discounted (d) effects (E) associated with this level of performance on the criterion test.

The weights are the relative values assigned by those people affected by where the criterion score is set. This would include a broad spectrum of people, including the individual himself or his legitimate representative. The effects are discounted in the customary way to balance off the advantages of immediate over deferred benefits. The effects themselves are variables of the type mentioned in connection with establishing task and corpus criterion scores.

It should be explicitly understood that neither the symbols, nor the preceding discussions, prescribed any particular method of measuring the factors in the model. I have been deliberately vague on these matters because they raise a question that is logically subsequent to the ones we are addressing here: namely, what should be the general content and form of the model? However, it should be noted that the results of this model can never be any more valid than the tests and measurements employed to apply it.

At least 4 criteria are relevant to the evaluation of this model. First, the model must be consistent with the values of our society. Second, it must take into account all of the major classes of variables that are relevant to the problem of identifying a criterion score. Third, it should be scientifically feasible to operationalize the model in a reasonably satisfactory manner. And fourth, the model should be practically feasible to apply. It seems clear to me that the model does meet all of these criteria, at least at a minimal level. However, I will leave its detailed examination to the reader.

A rejected model

Before we leave this topic, it seems important to examine briefly an alternative model that was explored and then rejected on ethical grounds. In recent years we have seen a marked increase in the study of the manpower needs of the United States and of the developing countries of the world. These investigations have dealt mainly with projecting the needs for personnel who are trained in the highly specialized skills involved in some occupation that is essential to the economies of those nations. But a few, particularly those addressed to the economic problems of developing nations, have also dealt with education in basic areas such as literacy. These studies attempted to build models for identifying how many people should be trained. These are models that are analogous to the corpus criterion model in which we sought to determine what portion of materials people should be able to read.

This type of model is ethically unacceptable for deciding how much reading instruction to give an individual. When it is suggested that we might like to know what proportion of people in our society should be literate, we are actually indulging in a euphemistic phrasing of the question of what proportion of the people in our society should be forced into illiteracy. And this implies that the policy makers have the right to decide who should learn to read and who should not. It is true, of course, that our educational resources are limited and that we may not be able to raise everyone to the level of literacy that we would ideally prefer. However, it seems unacceptable to allocate resources in this way. Suppose that it is partially true that the ability to read competently is an essential prerequisite to the exercise of our rights and responsibilities as citizens and to the participation in the social, economic, and cultural benefits of our society. And suppose that there is also some truth to the proposition that cultural advantage and disadvantage tend to perpetuate themselves. It seems that with this model we would be delegating to policy makers the right to create a caste system in our society.

Now, it remains true that our society and every other society can make the most of its resources by forecasting its future needs and by setting goals for meeting them. And we need information in order to do this. But that information is useless if it is cast in a form that is unacceptable to society. In our society we are willing to accept differences in the allocation of resources among people and differences in

people's eventual levels of attainment. But we prefer that those differences be explicitly determined by personal choice of the individual himself and by biological factors. Hence, we must reject this type of model.

A final comment

In essence, this article has been seeking a useful way to *ask* the question, *How well should a person learn to read?* We quickly rejected the arbitrary criteria previously used and then went on to reject models based on simplistic notions such as *more is better* and that *perfect mastery is ideal.* We also rejected partial solutions to this problem by recognizing that a person's literacy was jointly determined by both his reading ability and the readability of the materials that he needed to read. Instead, we chose to think in terms of models that regarded a person as literate when he could perform well enough to obtain the maximum value from the materials he needed to read. Consequently, we thereupon set out to examine, on the one hand, models that might tell us when a person was literate with respect to a single material or a corpus of materials and, on the other hand, models that might tell us when a person could read well enough to achieve his aspirations. Each of these models could probably be used with present techniques and produce modestly believable results, although each would be greatly improved if it received the benefit of further conceptual analysis and research. However, we must realize that these models, no matter how well they may be developed in the future, provide only preliminary and partial answers to the central question—*how well a person should learn to read,* given that literacy is jointly determined by reading ability and readability. The ultimate purpose of investigations of this sort is to help us make maximum use of our resources in realizing our goals, and this cannot be fully achieved until we have developed a model that permits us to *jointly* identify a criterion of literacy and readability.

REFERENCES

ALLEN, J. E. The right to read—target for the 70's. Address delivered before the National Association of State Boards of Education, September 23, 1969.

ANDERSON, RICHARD C. How to construct achievement tests to assess comprehension. *Review of Educational Research,* Spring 1972, 42, 145-170.

BART, WILLIAM M. A construction and validation of formal operational reasoning instruments. Paper read at the Amer-

ican Educational Research Association Convention, March 1970.

BETTS, EMMETT A. *Foundations of reading instruction*. New York: American Book, 1954.

BLOOM, BENJAMIN SAMUEL; ENGELHART, M. D.; FURST, J.; HILL, W. H.; & KROTWOHL, D. K. *Taxonomy of educational objectives handbook I: Cognitive domain*. New York: David McKay, 1956.

BLOOM, G. S. Learning for mastery. *Evaluation Comment*, March 1968, *1* (2), 1-8.

BORMUTH, J. R.; MANNING, J.; CARR, J.; & PEARSON, D. Children's comprehension of between and within sentence syntactic structures. *Journal of Educational Psychology*, October 1970, *61*, 349-357.

BORMUTH, JOHN R. *Implications and use of cloze procedure in the evaluation of instructional programs*. Los Angeles: University of California, (Occasional Report No. 3), Center for the Study of Evaluation Instructional Programs, 1967.

BORMUTH, JOHN R. An operational definition of comprehension instruction. In K. S. Goodman & J. F. Fleming (Eds.) *Psycholinguistics and the Teaching of Reading*. Newark: International Reading Association, 1969.(c)

BORMUTH, JOHN R. Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly*, Spring 1969, *4*, 358-367.(b)

BORMUTH, JOHN R. *Development of Readability analyses*. (Report of Project No. 7-0052) University of Chicago, 1969.(c)

BORMUTH, JOHN R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

BORMUTH, JOHN R. *Development of standards of readability*. (Report of Development Project No. 9-0237), University of Chicago, 1971.

BOND, GUY L. & TINKER, MILES A. *Reading difficulties: Their diagnosis and correction*. New York: Appleton-Century-Crofts, 1967.

CARROLL, JOHN B. A model of school learning: *Teachers College Record*, November 1963, *64*, 723-733.

DAVIS, FREDERICK B. *Educational measurements and their interpretation*. Belmont, California: Wadsworth, 1964.

FINN, PATRICK J. An algorithm for deriving operationally defined comprehension questions from written texts. Unpublished doctoral dissertation, University of Chicago, 1973.

GAGNE, ROBERT M. *The conditions of learning*. New York: Holt, Rinehart and Winston, 1965.

GIBSON, ELEANOR J. The ontogeny of reading. *American Psychologist*, February 1970, *25*, 136-143.

GOODMAN, KENNETH S. An analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly*, Fall 1969, *5*, 9-30.

HARRIS, ALBERT J. *Effective teaching of reading*. New York: David McKay, 1962.

HIVELY, W., II; PATTERSON, H. L.; & PAGE, SARA H. Generalizability of performance by job corps trainees on a universe-defined system of achievement tests in elementary mathematical calculation. Paper read at the American Educational Research Association Convention, February 1965.

LORGE, I. S. Readability formulae — an evaluation. *Elementary English*, February 1949, *36*, 86-95.

MAYO, S. F. Mastery learning and mastery testing. *NCME, Measurement in Education*, March 1970, *I*, 1-4.

MC GINITIE, W. H. Contextual constraint in English prose paragraphs. *Journal of Psychology*, January 1961, *51*, 121-130.

MILLER, D. M. Content, items, decisions: the orientation of curriculum assessment surveys to curriculum management and modification. *Educational Technology*, in press.

MOWRER, O. HOBART. The psychologist looks at language. *American Psychologist*, June 1954, *9*, 660-694.

OSGOOD, CHARLES E. On understanding and creating sentences. *American Psychologist*, March 1963, *18*, 735-751.

RANKIN, EARL F. Cloze procedure—a survey of research. *Yearbook of the South West Reading Conference*, 1965, *14*, 133-148.

ROTHKOPF, ERNST Z. Learning from written instructional materials: an exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, November 1966, *3*, 241-249.

TAYLOR, WILSON L. Application of "cloze" and entropy measures to the study of contextual constraint in samples of continuous prose. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, 1954.

THORNDIKE, E. L. Reading and reasoning: a study of mistakes in paragraph reading. *Journal of Educational Psychology*, October 1917, *8*, 323-332.

VENEZKY, RICHARD L. English orthography; its graphic structure and its relation to sound. *Reading Research Quarterly*, Spring 1967, *2*, 74-103.